MAP Growth Item Parameter Drift Study

January 2022

Wei He, NWEA Psychometric Solutions

Acknowledgments: This report benefited from the editorial assistance of Kelly Rivard.

Suggested citation: He, W. (2022). *MAP Growth item parameter drift study.* NWEA Research Report.

Table of Contents

List of Tables

List of Figures

Executive Summary

To ensure that student academic growth in a subject area is accurately captured, it is imperative that the underlying scale remains stable over time. As item parameter stability constitutes one of the factors that affects scale stability, NWEA® periodically conducts studies to check for the stability of the item parameter estimates for MAP® Growth™.

This report documents a routine item parameter drift study with its primary purpose to check for the parameter stability of MAP Growth items. Over the past decade, MAP Growth item calibration has adopted several changes in areas such as calibration sample selection and item calibration procedure to improve item calibration throughput and meet growing business needs. While these changes were implemented only after empirical studies showed that scale stability was not impacted, item parameter estimate stability has not necessarily been guaranteed at the individual item level.

The items of interest in the study were those calibrated before May 24, 2013, when iterative grade range (IGR) was used to establish the calibration sample for MAP Growth item calibration. The study started by re-estimating the items of interest with test events administered between Fall 2017 and Spring 2019 to understand the extent of the drift in item parameter estimates, followed by identifying items that were likely unstable using the Robust Z method. The final step was to conduct an impact analysis that examined the impact of the observed item parameter drift on ability estimates for students.

While the results suggest that between 4.2% and 6.9% of items across subjects were flagged as unstable by the Robust Z procedure and need further review to determine whether their parameters should be adjusted, the item-level and test-level analysis results indicate that the MAP Growth measurement scales are remaining stable.

1. Introduction

MAP® Growth™ is an adaptive interim assessment designed to measure achievement and growth in grades K–12 mathematics, reading, language usage, and science. Administered in the fall, winter, and spring, with an optional summer administration, MAP Growth reports scores on the Rasch Unit (RIT) vertical scale, which allows for the measurement of within- and between-year growth in student learning. Each subject has its own RIT scale.

Developing and maintaining vertical scales for educational achievement assessments is a complex yet challenging task. Aside from the psychometric issues, the educational ecosystem is constantly changing in nature. For example, the content students are learning may change noticeably from grade to grade, causing discontinuity in learning between grades. The order of the content that is taught, even to students in the same grade, may vary from school to school, causing item difficulty level to be systematically different in different school terms. Curriculum and instructional materials may also change from year to year at the same school. These changes create the challenge of measuring student achievement over time consistently yet accurately.

MAP Growth RIT scales were developed based on the Rasch item response theory (IRT) model that assumes item parameter invariance (i.e., the parameter value for the same item should not change systematically over multiple testing occasions). This invariance property is exceptionally valuable as it can provide capability to build measurement scales expected to maintain measurement characteristics even if test forms get changed later. In practice, however, IRT item parameter estimates will not be invariant. They are likely to drift upwards or downwards for different reasons, which, aside from what was mentioned beforehand, include item disclosure by previous students or change in the test construct over time. Additionally, when test items are reused often, they are likely to be overexposed, threatening item and test security. While computerized adaptive testing (CAT) commonly uses item exposure control methods to reduce the risk of item overexposure, item security is still widely acknowledged as a major problem in CAT due to its nature as a continuous test. If items are administered as often as daily, some items may become eventually known to new students after a while.

When item parameter invariance does not hold, item parameter drift occurs. The presence of item parameter drift can cause a series of consequences such as decreasing test validity, biased person ability estimates, misclassification of examinees, or scale drift. One way to detect item parameter drift as early as possible is to routinely re-estimate item parameters based on adaptive test data collected consecutively and compare the estimates with the values obtained in the initial calibration. If the difference is found to be significant, it is likely that item parameter drift has occurred, and that item may need to be removed or updated with the new parameter estimate. Examples of methods that are often used to detect item parameter drift in the Rasch model context include the Robust Z statistic (Huynh & Rawls, 2011) and “0.3 Logit Difference” (Miller et al., 2004).

The *Standards for Educational and Psychological Testing* recommend that “Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported” (AERA et al., 2014, p. 103, Standard 5.6). To ensure that student academic growth in a subject area is accurately captured, it is imperative that the underlying scale remains stable over time. As item parameter stability constitutes one of the factors that affects scale stability, item parameter estimate stability is usually examined to evaluate scale stability.

This report documents the results of an item parameter drift study designed to check the parameter stability of MAP Growth items. Over the past decade, MAP Growth item calibration has adopted several changes in areas such as calibration sample selection and item calibration procedure to improve item calibration throughput and meet growing business needs. The latest changes occurred with the adoption of the new item calibration tool known as Pychometrik in June 2021 that has removed the Pass 2 procedure (see Andrich et al., 2016), adopts the proportional curving fitting method to derive item difficulty, and uses the all grade calibration (AGC)[1] sample exclusively (He et al., 2021). While a series of empirical studies over the years have shown that scale stability has not been impacted by these changes, item parameter estimate stability has not necessarily been guaranteed at the individual item level. Studies such as this one will therefore need to be conducted on a regular basis to ensure the stability of parameter estimates for MAP Growth across time.

The items of interest in this study were those calibrated before May 24, 2013, when iterative grade range (IGR) was used to establish the calibration sample for MAP Growth item calibration. The study started by re-estimating the items of interest to understand the extent of the drift in item parameter estimates from item calibrations separated by at least four years, followed by identifying items that were likely unstable using the Robust Z method (Huynh & Rawls, 2009). The final step was to conduct an impact analysis that examined the impact of the observed item parameter drift on ability estimates for students. These items were re-calibrated using the most recent MAP Growth item calibration method implemented since June 2021 and test events administered between Fall 2017 and Spring 2019 (i.e., at least four years after the initial calibration).

[1] AGC identifies a calibration sample consisting of all students exposed to the same item. Its counterpart is called iterative grade range (IGR), which uses an iterative procedure to identify the best fitting grade(s) exposed to the same item and uses that subset of student responses to derive item parameter estimates.

2. Data and Method

2.1. Study Sample

The study included items successfully calibrated before May 24, 2013, but used in test events between Fall 2017 and Spring 2019. The original item list contained 12,063 unique items. As some of those items were retired or repurposed for another use such as sample items, the final list contained 9,800 items.

To apply the Robust Z procedure to detect item drift, the items needed to be re-estimated first. Babcock and Albano (2012) found that, in the context of credentialing a test that uses linear forms, a Rasch scale may remain stable for 15 ±3 years under conditions of little item parameter drift and small to moderate periodic changes in the latent trait, and substantial item parameter drift or large changes in the latent trait can dramatically reduce the longevity of the scale. As MAP Growth tests are used daily, we have decided to use test events collected between Fall 2017 and Spring 2019 to create the new item difficulty estimates (i.e., at least four years' time had elapsed between the initial item calibration and the test events used to create the new item difficulty estimates).

Table 2.1 presents the number and percentage of items included in the study by subject, as well as the summary descriptive statistics of the test events used to re-estimate the item difficulty. As the result of using item responses collected in two years for item recalibration (i.e., from 2017 to 2019), the average number of test events (i.e., calibration sample size) was massive, with reading items having the largest (278,864) and science items having the smallest (47,869). The use of a large calibration sample is always likely to produce stable item parameter estimates.

Table 2.1. Study Sample

Subject	Items		#Test Events (Fall 2017 – Spring 2019)			
	N	%	Mean	SD	Min.	Max.
Math	4,717	48.1	218,230	188,323	1,009	1,581,045
Reading	2,757	28.1	278,864	253,741	2,621	1,442,480
Language Usage	1,164	11.9	142,146	85,533	2,878	619,388
Science	1,162	11.9	47,869	37,331	2,086	183,256
Total	9,800	100.0	205,872	203,159	1,009	1,581,045

2.2. Robust Z Procedure

The Robust Z statistic (z_R; Huynh & Rawls, 2009), which originated from "robust statistical procedure," has been widely used to assess Rasch item difficulty stability in various large-scale assessments. A z-score, known as a z-value or standard score, is a measure of how many standard deviations below or above the population mean a raw score is. A z-score is calculated by $(X - \mu)/\sigma$, where X is an individual score, μ is the mean, and σ is the standard deviation. As both the mean and standard deviation can be influenced by outlying observations, the Robust Z score, which is much more tolerant of outliers, is recommended for use to remedy this problem. With the Robust Z score, outliers can be detected reliably even in the presence of outliers in the data used to compute the median and median absolute deviation.

To compute the Robust Z statistic (z_R), the mean is replaced by the median (denoted by Md) and the standard deviation is replaced by 0.74 × the interquartile range (denoted by IQR) to match the standard deviation of the normal distribution. That is:

$$z_R = \frac{D - Md}{0.74 \times IQR}$$

In the context of detecting item difficulty stability, D indicates the difference in item difficulties obtained from two separate calibrations, and Md and IQR indicate the median and the interquartile range for the difficulty differences for the items of interest.

The Robust Z score follows (asymptotically) the standard normal distribution with zero mean and unit standard deviation. As such, a level of significance (two-tailed a) may be selected and a positive critical value z^x may be set. Items with z_R smaller than the z^x in absolute value can be declared as "stable," and items with z_R larger than the z^x in absolute value can be declared as "unstable."

2.3. Analysis

The following steps were undertaken for this study:

Step 1: Re-estimate the item parameters using the new MAP Growth item calibration procedure implemented since June 2021 and test events collected between Fall 2017 and Spring 2019. Exclude items with less than 1,000 test events from the analysis.

Step 2: Examine the difference between the original and new item parameter estimates:

a. Compute the average difference in difficulty estimate ($\hat{b}_{new} - \hat{b}_{original}$) and the correlation coefficients between the original and new item parameter estimates.
b. Allocate items into one of the following difficulty parameter difference categories (in logit) based on $\hat{b}_{new} - \hat{b}_{original}$. Aggregate the results by subject. 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

Step 3: Apply the Robust Z procedure to identify unstable items in different subjects. The tests were two-sided with significance level as 0.05.

Step 4: Examine the impact of the observed item parameter drift on students' scores:

a. Randomly select 3,000 test events administered between Fall 2011 and Spring 2013 in each subject and term and rescore them using the same operational ability estimate procedure as MAP Growth is currently using but with the new item difficulty estimates obtained in Step 1. The rescoring will yield a new score for each test event.
b. Compare the new ($\hat{\theta}_{new}$) and original ($\hat{\theta}_{original}$) scores from each individual student across subjects and terms to identify score differences caused by the change in item parameter estimates.

3. Results

3.1. New Item Parameter Estimates

Table 3.1 presents the summary statistics of the original ($\hat{b}_{original}$) and new ($\hat{b}_{new}$) item difficulty estimates for each subject, along with the difference between them (Δ). As items with less than 1,000 responses were excluded from the analysis, the number of items used in the analysis was slightly less than that in Table 2.1. The average change in item difficulty estimates was 0.03, -0.03, 0.06, and 0.02 logit in math, reading, language usage, and science, respectively. These changes were slightly larger, if not the same, than what is reported in Kingsbury (2003) that the average change was -0.011 and -0.017 logit in math and reading, respectively. These changes can be considered quite small, despite that some individual items had large differences (e.g., the maximum difficulty difference for math items was 1.35 logit). The distributions of the item difficulty difference, as shown in Figure 3.1, are normal and symmetrical around zero for all subjects.

Table 3.1. Summary Statistics of the Original and New Item Difficulty Estimates and their Differences

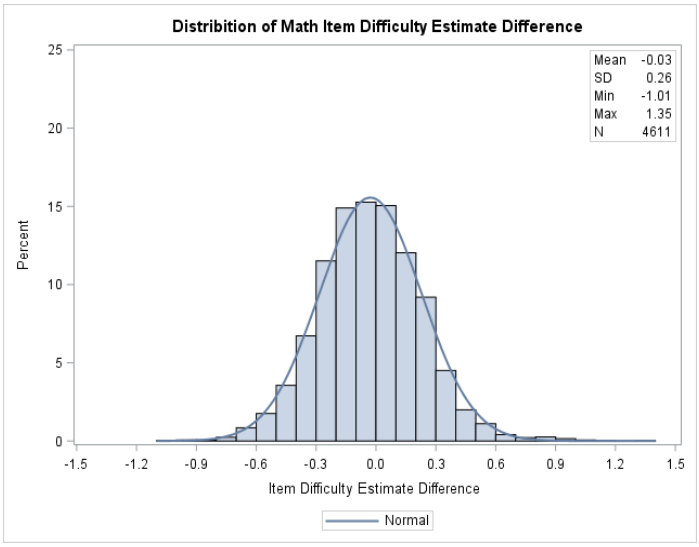
Subject	#Items	Mean			SD			Min.			Max.		
		$\hat{b}_{original}$	$\hat{b}_{new}$	Δ	$\hat{b}_{original}$	$\hat{b}_{new}$	Δ	$\hat{b}_{original}$	$\hat{b}_{new}$	Δ	$\hat{b}_{original}$	$\hat{b}_{new}$	Δ
Math	4,611	0.82	0.79	-0.03	3.34	3.35	0.26	-8.20	-8.27	-1.01	8.40	8.69	1.35
Reading	2,743	-1.40	-1.43	-0.03	3.08	3.11	0.21	-8.50	-8.50	-1.04	5.70	5.93	0.83
Language Usage	1,155	0.30	0.36	0.06	1.80	1.75	0.22	-4.70	-4.37	-0.87	5.20	5.62	1.18
Science	1,162	0.18	0.21	0.02	1.62	1.61	0.19	-5.20	-5.07	-0.67	4.40	4.57	0.76

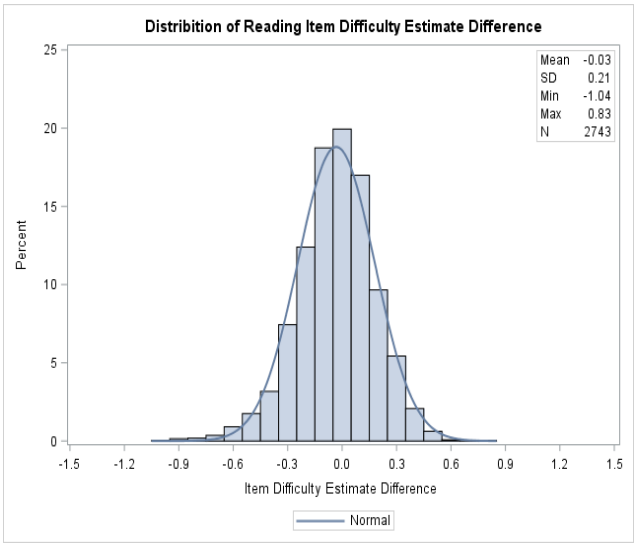
Note. $\boldsymbol{\Delta = \hat{b}_{new} - \hat{b}_{original}}$

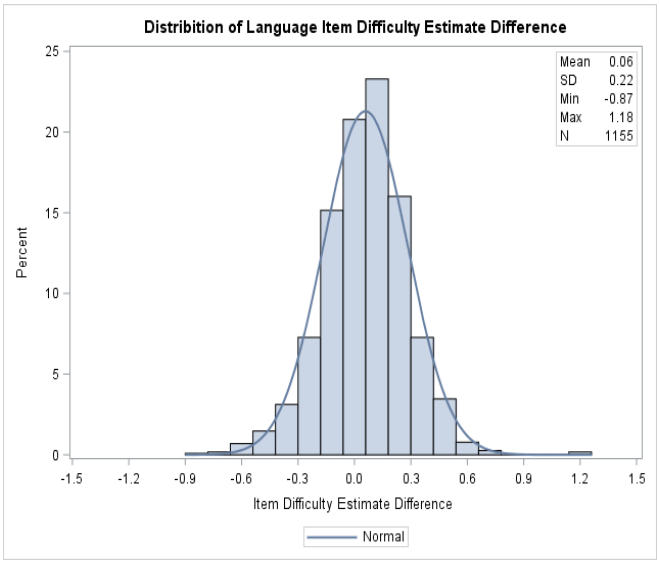
Figure 3.2 presents the scatterplots between the original and new difficulty estimates in each subject, along with their observed correlations at the low right corner in each figure. The correlations were extremely high, above 0.99 for all four subjects.

Figure 3.3 presents the percentage of items falling into the different difficulty parameter difference categories by subject based on $\hat{b}_{new} - \hat{b}_{original}$. The blue and red bars, which indicate the "0.3 Logit Difference" (Miller et al., 2004) categories, show that at least 80% of items have differences within 0.3 logits for almost all subjects except math that came a bit short of 80%. In large-scale assessment programs using the Rasch model, the "0.3 Logit Difference" is widely used as a rule of thumb to flag 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.

Figure 3.1. Distributions of Item Difficulty Estimate Differences







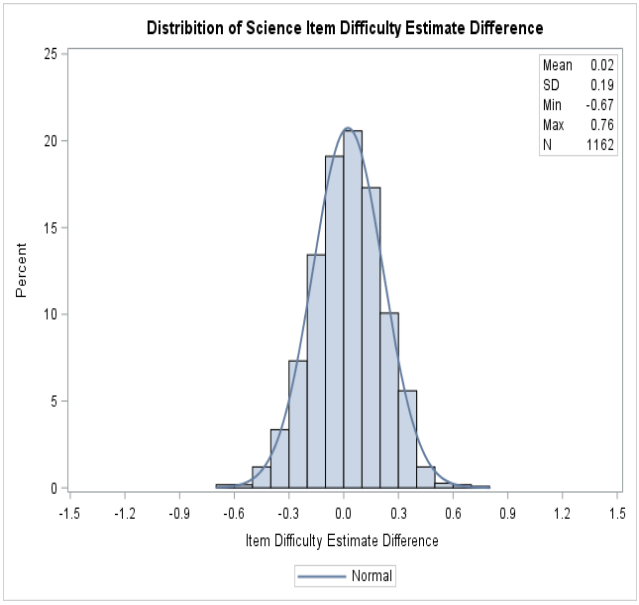


Figure 3.2. Scatterplots between the Original and New Difficulty Estimates

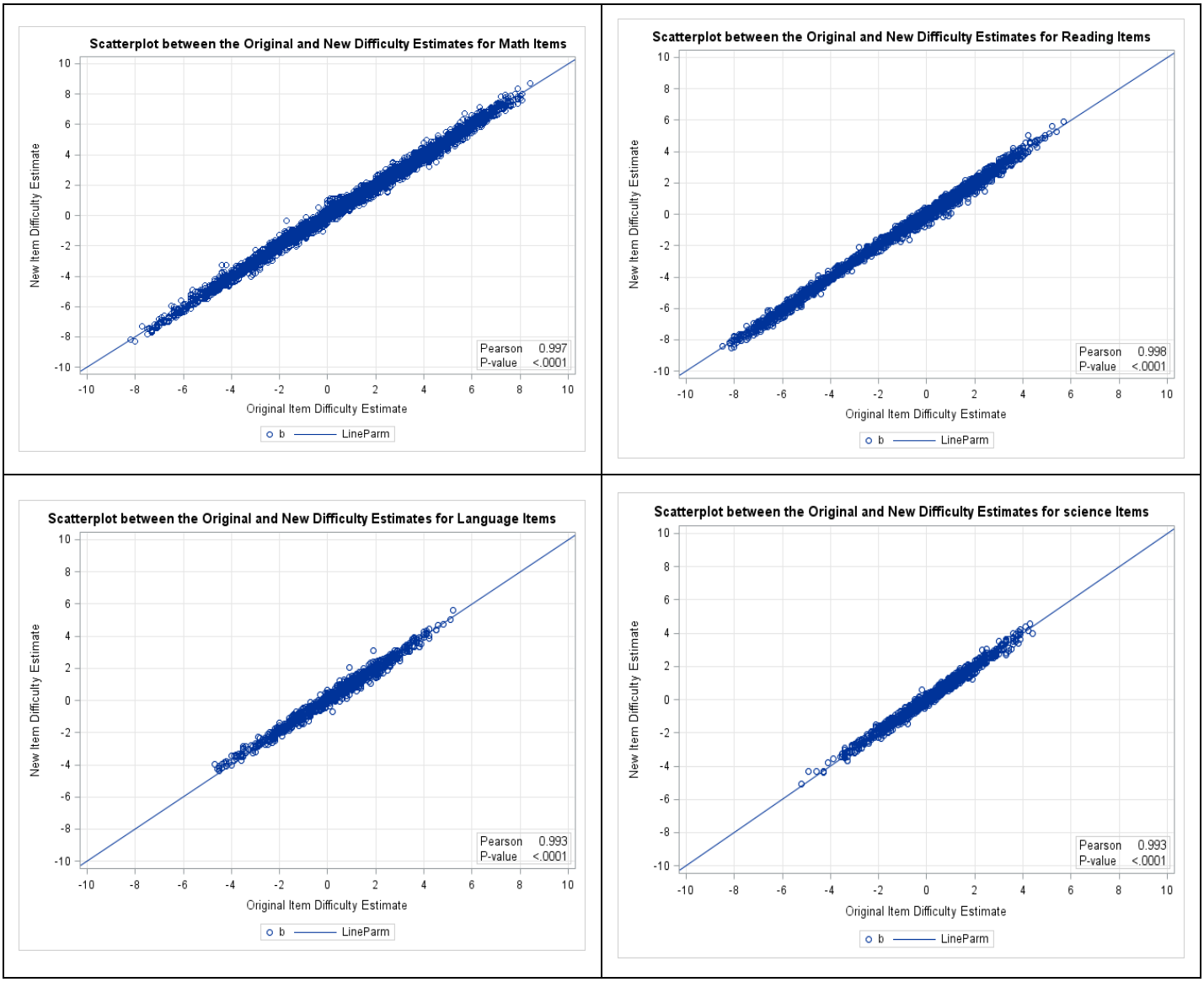
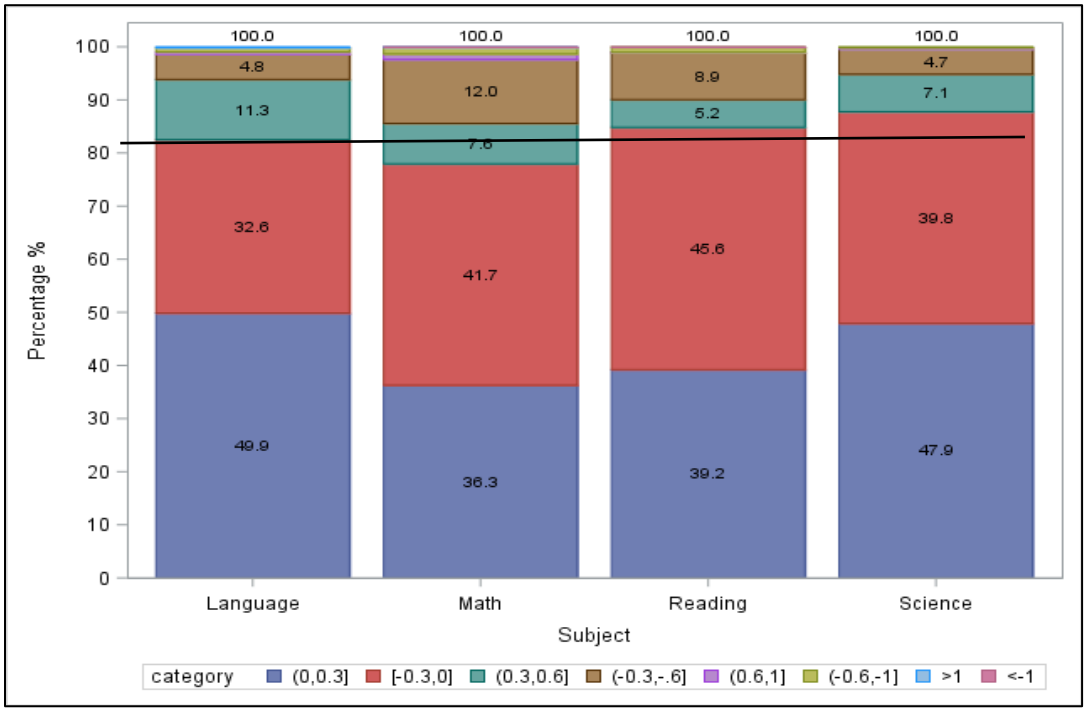


Figure 3.3. Percentage of Items in Different Difficulty Parameter Difference Categories



3.2. Robust Z Statistics Results

As shown in Table 3.2, between 4.2% and 6.9% of items across subjects were flagged as unstable by the Robust Z procedure. Those percentages, except for science items, are slightly larger than the 5% nominal error rate used in the Robust Z procedure, suggesting that some amount of item parameter drift was occurring in the items associated with each measurement scale. In general, the average difficulty estimate difference of the unstable items was slightly larger than that of the stable items across all subjects, though the magnitudes of the average differences ranged from 0.02 to 0.19 logit.

Figure 3.4 presents the distributions of difficulty estimate differences ($\hat{b}_{new} - \hat{b}_{original}$) of the stable and unstable items across subjects. The figures unanimously show that items with differences in parameter estimates within 0.3 logit fall into the "stable" category, consistent with the "0.3 Logit Difference" rule. The magnitude of the cut-off value to flag items into the stable and unstable categories vary between 0.34 and 0.51 logit across subjects. Figure 3.5 plots the original item difficulty distributions ($\hat{b}_{original}$) for the stable and unstable items across subjects. These plots indicate 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.

Table 3.2. Number of Flagged and Unflagged Items by Robust Z Procedure and their Summary Descriptive Statistics

Subject	#Items	Robust Z			$\widehat{b}_{new}$				$\widehat{b}_{original}$				$\widehat{b}_{new} - \widehat{b}_{original}$			
		Status	#Items	%	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Math	4,611	Stable	4,372	94.8	0.76	3.36	-8.27	8.69	0.79	3.35	-8.20	8.40	-0.03	0.22	-0.52	0.45
		Unstable	239	5.2	1.36	3.08	-5.93	7.95	1.30	3.16	-6.50	7.40	0.06	0.64	-1.01	1.35
Reading	2,743	Stable	2,553	93.1	-1.45	3.08	-8.50	5.93	-1.43	3.06	-8.50	5.70	-0.02	0.17	-0.41	0.36
		Unstable	190	6.9	-1.23	3.48	-8.44	5.59	-1.04	3.37	-8.00	5.20	-0.19	0.49	-1.04	0.83
Language Usage	1,155	Stable	1,075	93.1	0.34	1.75	-4.37	5.62	0.27	1.79	-4.60	5.20	0.07	0.18	-0.34	0.47
		Unstable	80	6.9	0.64	1.65	-3.99	3.53	0.68	1.89	-4.70	3.90	-0.04	0.55	-0.87	1.18
Science	1,162	Stable	1,113	95.8	0.20	1.59	-5.07	4.57	0.17	1.60	-5.20	4.30	0.03	0.17	-0.35	0.40
		Unstable	42	4.2	0.33	1.98	-4.30	3.97	0.45	2.09	-4.90	4.40	-0.12	0.45	-0.67	0.76

Figure 3.4. Histograms of Difficulty Estimate Differences for Stable and Unstable Items

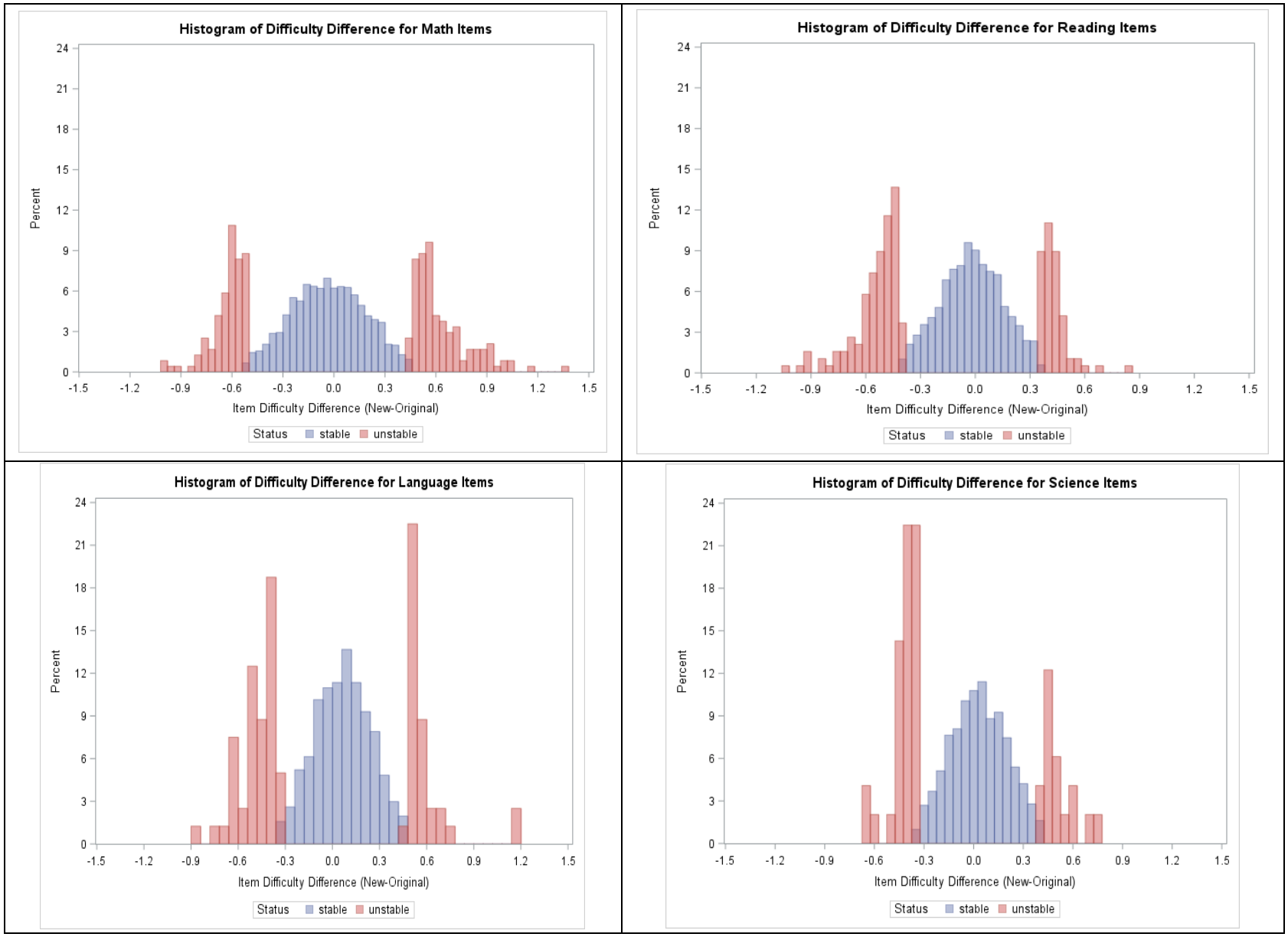
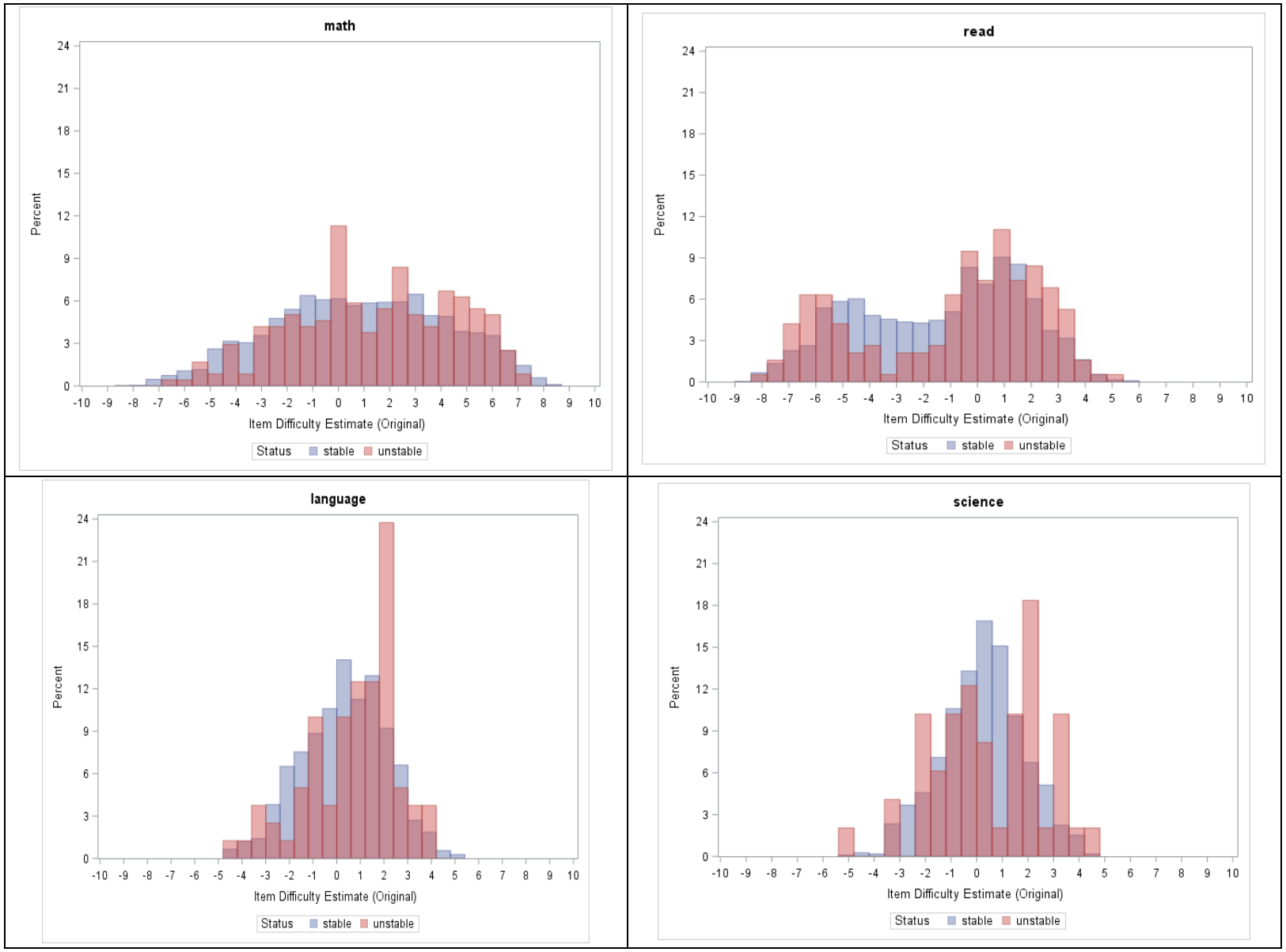


Figure 3.5. Histograms of Original Difficulty Estimates for Stable and Unstable Items



3.3. Impact of the Observed Parameter Drift on Ability Estimate

The impact of the observed parameter drift on ability estimate was examined by comparing the original and new ability estimates that were obtained via rescoring using new item parameter estimates and the old item responses. A total of 3,000 test events were randomly selected in each subject and term from Fall 2011 to Spring 2013.

Table 3.3 reports the summary statistics of the average differences between the original and new ability estimates ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{new}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{original}}$). The average ability estimate differences for all subjects and terms is quite small, with the largest magnitude being 0.03 logit (i.e., 0.3 RIT) observed for the Spring 2013 math tests. Considering that the MAP Growth RIT score is reported at an increment of 1 (i.e., 0.1 logit), the magnitudes of the average difference can be viewed as negligible. As shown in the maximum (i.e., Max.) column, the magnitudes of the largest difference between the new and original score at the individual test level across all terms of interest were 0.22 logit for math, 0.29 logit for reading, 0.33 logit for language usage, and 0.28 logit for science.

The adaptive nature of MAP Growth implies that the number of items with new parameter estimates encountered by each individual student varies. Table 3.3 also reports the summary statistics of the number of items with difficulty estimates different from their original ones that were encountered in tests across subjects and terms (#Items with $\widehat{\boldsymbol{b}}_{\boldsymbol{new}}$). The minimum number is as low as 1, but the maximum number can be just a few items short of the full test length. For example, among all Spring 2013 math tests, the maximum number of items with parameter estimates different from their original ones was 36, just 15 items less than the full-length test. Furthermore, across terms, the average number of items with difficulty estimates different from their original ones ranged between 9 and 19 for math, 5 and 11 for reading, 6 and 10 for language usage, and 3 and 8 for science, equivalent to 18~38% for math, 13~27% for reading, 13~20% for language usage, and 11~25% for science in terms of the percentage over the entire test length (i.e., the last column in Table 3.3).

Overall, the new and original ability estimates were almost perfectly correlated, above 0.99, for all subjects and terms.

Table 3.3. Summary Descriptive Statistics of $\widehat{\boldsymbol{\theta}}_{\boldsymbol{new}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{original}}$ and Number of Items with $\widehat{\boldsymbol{b}}_{\boldsymbol{new}}$

Term	Subject	#Test Events	$\widehat{\boldsymbol{\theta}}_{\boldsymbol{new}} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{original}}$				#Items with $\widehat{\boldsymbol{b}}_{\boldsymbol{new}}$				Test Length	%Average #Items with $\widehat{\boldsymbol{b}}_{\boldsymbol{new}}$ /TL
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.		
Spring 2013	Math	3,000	-0.03	0.07	-0.50	0.15	19	5	6	36	51	38
	Reading	2,999	-0.02	0.06	-0.31	0.12	11	4	1	27	42	27
	Language Usage	2,996	0.01	0.02	-0.07	0.16	10	3	1	26	51	20
	Science	3,000	0.00	0.03	-0.12	0.14	8	3	1	19	30	25
Winter 2013	Math	2,998	-0.02	0.06	-0.43	0.18	16	5	4	32	50	33
	Reading	2,996	-0.02	0.06	-0.24	0.18	9	4	1	21	42	22
	Language Usage	2,998	0.01	0.03	-0.12	0.33	10	4	1	35	50	19
	Science	3,000	0.00	0.03	-0.16	0.24	7	3	1	20	30	24
Fall 2012	Math	3,000	-0.01	0.04	-0.31	0.21	15	5	2	37	50	29
	Reading	2,964	-0.01	0.05	-0.46	0.29	8	3	1	21	42	20
	Language	3,000	0.01	0.04	-0.11	0.28	9	4	1	27	50	18
	Science	2,998	0.01	0.04	-0.09	0.27	6	3	1	15	30	20

Term	Subject	#Test Events	$\widehat{\theta}_{new} - \widehat{\theta}_{original}$				#Items with $\widehat{b}_{new}$				Test Length	%Average #Items with $\widehat{b}_{new}$ /TL
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.		
Spring 2012	Math	2,998	-0.01	0.03	-0.28	0.22	9	5	1	28	50	18
	Reading	2,999	-0.01	0.04	-0.21	0.25	6	3	1	20	42	14
	Language Usage	2,998	0.01	0.03	-0.08	0.26	7	4	1	29	50	15
	Science	2,999	0.01	0.02	-0.07	0.24	4	2	1	15	30	14
Winter 2012	Math	2,476	0.00	0.04	-0.17	0.18	15	5	1	33	50	30
	Reading	2,554	0.00	0.03	-0.11	0.21	6	3	1	17	40	16
	Language Usage	3,000	0.01	0.03	-0.07	0.26	7	3	1	19	50	14
	Science	3,000	0.00	0.03	-0.14	0.24	4	2	1	13	30	12
Fall 2011	Math	2,613	0.00	0.03	-0.14	0.17	12	5	1	35	50	24
	Reading	2,588	0.01	0.03	-0.13	0.27	5	3	1	17	40	13
	Language Usage	3,000	0.01	0.04	-0.14	0.29	6	3	1	24	50	13
	Science	3,000	0.00	0.04	-0.16	0.28	3	2	1	14	30	11

4. Conclusion and Discussion

While the primary focus of this study was to identify MAP Growth items that might be drifting in difficulty, the study took a further step to examine the impact of parameter drift on student scores, assuming that the difficulty estimates of some items have changed over time. The results of the item-level analysis (i.e., the item parameter drift study) suggest that the average differences in difficulty estimates of the items of interest are negligible in all four subjects. For a large proportion of items, the difficulty estimates derived from the test events collected at least four years after they were initially calibrated remain consistent with their original ones. Depending on the subject, between 4.2% and 6.9% of items were flagged as unstable by the Robust Z procedure and will be further reviewed to determine whether their parameters should be adjusted.

The analysis examining the impact of parameter drift on student scores can be viewed at the test level. This analysis accepts that individual item parameters may change but asks whether the changes affect student scores and the decisions made with scores. It is clear from the results that, on average, the impact of parameter changes on ability estimates was negligible for all four subjects. However, the magnitude of the impact varies across individual students. This can be attributed to the adaptive nature of MAP Growth and the large MAP Growth item pool. Students see a very small proportion of items out of the entire item pool in an adaptive test, so it is likely that the proportion of items drifting in difficulty but administered in each individual student's test varies. Some might see more, and some might see fewer. Depending on the magnitude of the drift, the impact on student scores will vary as well.

The item calibration procedure used in this study is slightly different from that used to derive the original difficulty parameters, but the results further confirm that both the old and new item calibration procedures yield comparable parameter estimates. However, it is also likely that this change, intertwined with other factors, played a role in parameter estimate drift for items flagged as unstable. Nevertheless, both the item-level and test-level analysis results indicate that the MAP Growth measurement scales are remaining stable. This conclusion is consistent with what past studies (e.g., Kingsbury, 2003) have found.

5. References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.

Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement*, *76*(3), 412–435.

Babcock, B., & Albano, A. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*, 549–564. doi: 10.1177/0146621612455090.

He, W., Bo, E., Meyer, J. P., & Grandgeorge, R. (2021). *A comparison of item parameter estimates in Pychometrik and the existing item calibration tool*. NWEA. https://www.nwea.org/content/uploads/2021/05/Pychometrik-Item-Parameter-Estimate-Comparability-Study-2021-01-15_NWEA_report.pdf

Kingsbury, G. G. (2003). *A long-term study of the stability of item parameter estimates*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Miller, G. G., Rotou, O., & Twing, J. S. (2004). Evaluation of the 0.3 logits screening criterion in common item equating. *Journal of Applied Measurement*, *5*(2), 172–177.