MAP Growth K–2 Item Fit Analysis: A Follow-up Study

May 2020

Wei He, Ph.D., NWEA Psychometric Solutions

Suggested citation: He, W. (2020). *MAP Growth K–2 item fit analysis: A follow-up study*. NWEA.

Table of Contents

List of Tables

List of Figures

1. Introduction

Item fit analysis examines how accurately observed response data fit the underlying model. In test analysis, item fit can be used to validate the calibration process of item parameters. The purpose of this study is to examine the fit of MAP® Growth™ K–2 items involved in the 2020 scale alignment study to realign the scales underlying the MAP Growth K–2 and 2–5 Reading and Mathematics tests (Thum & Kuhfeld, 2020). To make sure the items with the adjusted RIT values still fit the underlying Rasch model, this study examines the model-data fit of these items using two different samples (i.e., AllYr and LatestYr). For each sample, infit and outfit indices, along with point measure correlations were calculated, followed by a comparison of how these indices differ using the old and new (i.e., original and adjusted) item difficulties and person ability estimates. The results from the different samples were also compared with each other.

An item fit analysis study was conducted in May 2019 (He, 2019) based on the 2019 scale alignment study (Thum & Kuhfeld, 2019). This is a follow-up analysis to evaluate the fit of the items whose item difficulties were adjusted in the more recent 2020 follow-up scale alignment study (Thum & Kuhfeld, 2020). This study looked at different items than what were used in the 2019 analysis.

1.1. Data

Items of interest in this study were MAP Growth K–2 Reading and Mathematics items whose difficulties were adjusted in the 2020 scale alignment study. Responses to these items were from the MAP Growth K–2 test events administered between Fall 2015 and Fall 2020. To investigate the degree to which item fit can be affected by different samples, the following different samples were created in the same manner as in the 2019 item fit analysis (He, 2019) based on the same test events data used in the 2020 scale alignment study:

1. AllYr: Item responses from all the years in which an item was exposed between Fall 2015 and Fall 2020
2. LatestYr: Item responses between Fall 2019 and Fall 2020

Student RIT scores in these test events were adjusted in this follow-up item fit analysis using the following equations for Reading and Mathematics, respectively (Thum & Kuhfeld, 2019). Items with less than 300 responses were excluded from the study, resulting in a total of 285 Reading and 305 Mathematics items in the AllYr sample and 282 Reading and 301 Mathematics items in the LatestYr sample.

$$\text{Reading\_RIT}_{\text{Adjusted}} = 8.6874+0.9211\times\text{Reading\_RIT}_{\text{Old}} \quad (1)$$

$$\text{Math\_RIT}_{\text{Adjusted}} = 26.52+0.8314\times\text{Math\_RIT}_{\text{Old}} \quad (2)$$

Table 1.1 presents the descriptive statistics of difficulties for these MAP Growth K–2 items (i.e., the item RIT values), which are the same for both samples. For both content areas, the average new item RITs were slightly smaller than the average old item RITs, with the differences being 4 and 3 RITs for Reading and Mathematics, respectively. For both samples, the correlations between the old and new item RITs were 0.995 for both content areas.

Table 1.1. Descriptive Statistics of Item Difficulties

	#Items			RIT			
Content Area	AllYr	LatestYr	Item RIT Group	Mean	SD	Min.	Max.
Reading	285	282	New	175	17	125	208
			Old	179	18	120	215
Mathematics	305	301	New	171	27	123	236
			Old	174	32	117	253

As shown in Table 1.2, the average number of responses for Reading and Mathematics items in the AllYr sample are 53,876 and 50,879, respectively, and the average number of responses for Reading and Mathematics items in the LatestYr sample are 30,457 and 35,843, respectively.

Table 1.2. Descriptive Statistics of Item Responses

			Response Count per Item			
Sample	Content Area	#Items	Mean	SD	Min.	Max.
AllYr	Reading	285	53,876	54,139	576	590,006
	Mathematics	305	50,879	53,222	375	369,916
LatestYr	Reading	282	30,457	26,809	336	210,583
	Mathematics	301	35,843	33,049	326	196,573

1.2. Analysis Method

MAP Growth assessments operate on the Rasch model, and the most commonly used statistics to assess item fit for the Rasch model are infit and outfit. In a Rasch context, these statistics tell how accurately or predictably data fit the model. Infit, outfit, and point measure correlation used in this study are defined in Equations 3–5:

$$Infit_i = \frac{\sum_n^N (O_{ni}-P_{ni})^2}{\sum_n^N P_{ni}(1-P_{ni})} \tag{3}$$

$$Outfit_i = \frac{\sum_n^N \frac{(O_{ni}-P_{ni})^2}{P_{ni}(1-P_{ni})}}{N} \tag{4}$$

$$r_{pm_i} = \frac{\sum_{n=1}^N (O_{ni}-\bar{O})\left(\hat{\theta}_{ni}-\bar{\bar{\theta}}\right)}{\sqrt{\sum_{n=1}^N (O_{ni}-\bar{O})^2 \sum_{n=1}^N \left(\hat{\theta}_n-\bar{\bar{\theta}}\right)^2}} \tag{5}$$

where:

- O_{ni} is the observed response (either correct or incorrect) by examinee *n* to item *i*.
- P_{ni} is the probability of correct response based on the Rasch model that is calculated by $P_{ni} = \frac{1}{1+\exp\,(b_i-\hat{\theta}_n)}$, where b_i = item difficulty.
- $\hat{\theta}_n$ is the ability estimate for examinee *n*.
- $\bar{O}$ is the proportion correct for item *i*.
- $\hat{\theta}_{ni}$~~$\widehat{\theta_{ni}}$~~ is the ability estimate of examinee *n* who was administered item *i*.
- $\bar{\bar{\theta}}$ is the average ability estimate for examinees who were administered item *i*.
- r_{pm_i} is the point measure correlation for item *i*.

To examine item fit, the following analyses were conducted for each sample (i.e., AllYr and LatestYr) using SAS 9.4:

Step 1. Calculate the infit, outfit, and point measure correlations using Equations 3–5. For each item, two sets of values were calculated for each of these indices using each sample. One set was based on the old values (i.e., the original item difficulties, ability estimates, and item responses), and the other was based on the new values (i.e., the adjusted item difficulties, new ability estimates, and item responses). As mentioned earlier, the new ability estimates were obtained by applying Equations 1 and 2 depending on the content area.

Step 2. Calculate the distances between the infit and outfit statistics of each item to 1.0 and compare the differences based on the new and the old values according to Equations 6 and 7. The reason for doing so is that the expected values for both infit and outfit statistics are 1.0. The closer the values are to 1.0, the better the item fit.

$$Infit_{diff_i} = Abs\left(Infit_{new_i} - 1\right) - Abs\left(Infit_{old_i} - 1\right) \tag{6}$$

$$Outfit_{diff_i} = Abs\left(Outfit_{new_i} - 1\right) - Abs\left(Outfit_{old_i} - 1\right) \tag{7}$$

Step 3. Examine the point measure correlations. Items with a value less than 0.2 are flagged as poor-quality items.

Step 4. Flag the remaining items from Step 3 for potential misfit based on the new statistics using the following two sets of criteria: strong and weak. Both sets of items did not have any items in common with those from Step 3, but the “weak” set of items is a subset of the “strong” set of items. In other words, the strong and weak criteria were not applied to any of the items already flagged in Step 3, and items flagged based on the strong criteria could also be flagged based on the weak criteria. The purpose of using these two criteria is to compare how many items are flagged based on stringent vs. lenient criteria.

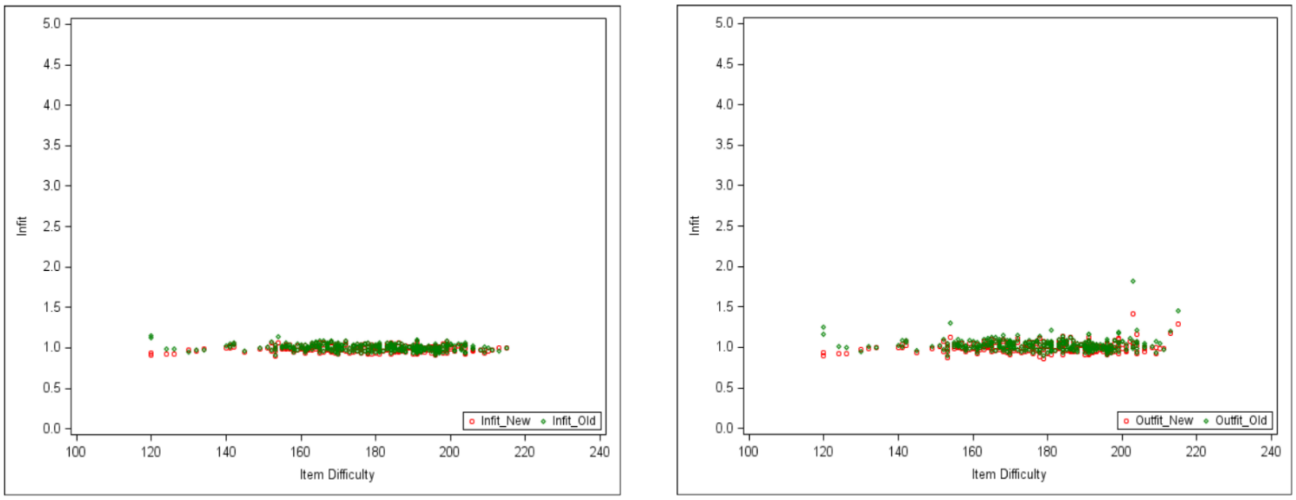
a. Strong: Flag items with new infit or outfit greater than 1.2 or less than 0.8.
b. Weak: Flag items with new infit or outfit greater than 1.5 or less than 0.5.

For items flagged for potential misfit based on the criteria in Step 4 in each sample, plot their item characteristics curves (ICCs) using the adjusted item difficulty and the observed proportion correct conditional on the new person ability estimates. These items will receive both content and psychometric reviews before being deactivated.

2. Results

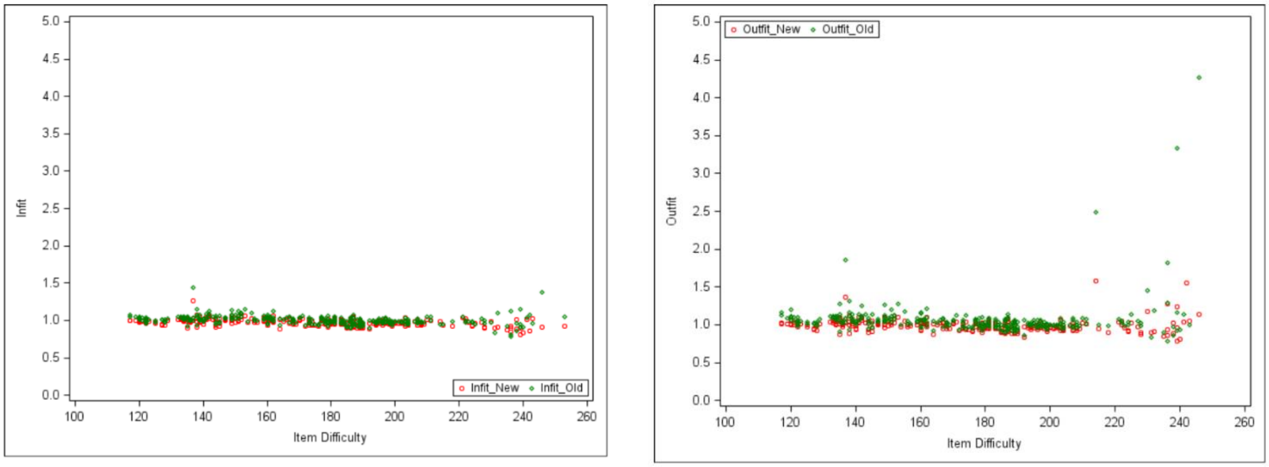
Figure 2.1 and Figure 2.2 plot both the new and old infit and outfit statistics against item difficulties for Reading and Mathematics using the AllYr samples[1]. Two Mathematics items in the AllYr sample with an old outfit value > 6 were excluded from Figure 2.2. As shown in the figures, the old and the new infit statistics do not appear to differ in a noticeable manner. However, for Mathematics, the new infit statistics for those difficult items (i.e., item RIT>220) appear to have been brought closer to 1 with the adjusted item difficulties. The new outfit statistics are closer to 1 compared to their old counterparts for those difficult items in both content areas. This suggests that the adjustment of item difficulties has helped with the improvement of the item fit. Compared with the Reading items, Mathematics items exhibited more variations in both the infit and outfit statistics. The outfit statistics showed more variations than the infit statistics for both content areas.

Figure 2.1. New and Old Infit and Outfit Statistics vs. Item Difficulty (AllYr Sample)—Reading



Infit **Outfit**

Figure 2.2. New and Old Infit and Outfit Statistics vs. Item Difficulty (AllYr Sample)—Mathematics



Infit **Outfit**

[1] The same finding was observed regardless of the sample, so only the scatterplots from the AllYr sample are presented in the report.

Table 2.1 presents the Pearson correlations between the new and old infit statistics. For Reading, the correlations are between the mid to high 0.8-0.9 range, suggesting that the new and the old fit statistics are highly correlated. For Mathematics, the correlations between the new and the old infit statistics for both samples are around 0.7 (i.e., moderately correlated), whereas the correlations between the new and the old outfit statistics were as high as 0.99.

Table 2.1. Pearson Correlation Coefficients between the New and the Old Fit Statistics

	Reading		Mathematics	
Sample	**Infit$_{New}$,Infit$_{Old}$**	**Outfit$_{New}$,Outfit$_{Old}$**	**Infit$_{New}$,Infit$_{Old}$**	**Outfit$_{New}$,Outfit$_{Old}$**
AllYr	0.85	0.88	0.72	0.99
LatestYr	0.85	0.87	0.68	0.99

Table 2.2 presents the summary infit and outfit item statistics and point measure correlations, including the mean, standard deviation (SD), and minimum and maximum infit and outfit statistics for the two samples. The results from this table echo the observations above that the fit of the items has been improved with the adjusted item difficulties from the scale alignment study. The same findings were observed for both Reading and Mathematics.

The distributions of both infit and outfit statistics are centered around 1.0. The closer the values are to 1.0, the better the item fit. As such, the distances of infit and outfit statistics from the expected value of 1.0 were computed and the differences were compared based on the new and the old values. For infit statistics, the average distances from the expected value of 1.0 for the old (i.e., *abs(Infit$_{old}$-1)*) and new statistics (i.e., *abs(Infit$_{new}$-1)*) are the same for Reading (i.e., 0.03) and for Mathematics (i.e., 0.04). However, for outfit statistics, the average distances from the expected value of 1.0 for the old (i.e., *abs(Outfit$_{old}$-1)*) and new statistics (i.e., *abs(Outfit$_{new}$-1)*) are 0.05 and 0.04 for Reading and 0.25 and 0.07 for Mathematics. That is, the absolute differences between the new statistics and 1.0 are smaller than those between the old statistics and 1.0, suggesting that the adjusted item difficulties have improved the model-data fit for both Reading and Mathematics items. The old and new statistics for the point measure correlations remained the same for both Reading and Mathematics items in the follow-up study, which is expected as the new person ability estimates for each content area were obtained by applying a linear equation to the old person ability estimates.

Table 2.2. Summary Infit and Outfit Statistics

		Reading				Mathematics			
Sample	**New and Old Infit and Outfit Statistics**	**Mean**	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
AllYr	$Abs(Infit_{new}-1)$	0.03	0.02	0.00	0.10	0.04	0.04	0.00	0.27
	$Abs(Infit_{old}-1)$	0.03	0.02	0.00	0.15	0.04	0.04	0.00	0.44
	$Abs(Outfit_{new}-1)$	0.04	0.04	0.00	0.42	0.07	0.40	0.00	6.99
	$Abs(Outfit_{old}-1)$	0.05	0.07	0.00	0.82	0.25	2.38	0.00	40.78
	$Infit_{diff}$	0.00	0.02	-0.09	0.07	0.00	0.04	-0.29	0.14
	$Outfit_{diff}$	-0.01	0.04	-0.41	0.08	-0.17	1.98	-33.79	0.14
	$Infit_{new}$	0.99	0.04	0.90	1.08	0.97	0.05	0.81	1.27
	$Infit_{old}$	1.00	0.04	0.90	1.15	1.00	0.06	0.78	1.44
	$Outfit_{new}$	1.01	0.06	0.86	1.42	1.01	0.41	0.78	7.99
	$Outfit_{old}$	1.03	0.08	0.90	1.82	1.22	2.38	0.78	41.78
	r_{pm_new}	0.31	0.06	0.17	0.48	0.31	0.08	0.06	0.57
	r_{pm_old}	0.31	0.06	0.17	0.48	0.31	0.08	0.06	0.57

Sample	New and Old Infit and Outfit Statistics	Reading				Mathematics			
		Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
LatestYr	$Abs(Infit_{new}-1)$	0.03	0.02	0.00	0.12	0.04	0.03	0.00	0.25
	$Abs(Infit_{old}-1)$	0.03	0.02	0.00	0.16	0.04	0.04	0.00	0.46
	$Abs(Outfit_{new}-1)$	0.04	0.04	0.00	0.34	0.07	0.40	0.00	6.99
	$Abs(Outfit_{old}-1)$	0.05	0.06	0.00	0.53	0.25	2.39	0.00	40.78
	$Infit_{diff}$	0.00	0.01	-0.09	0.06	0.00	0.04	-0.29	0.10
	$Outfit_{diff}$	-0.01	0.03	-0.26	0.07	-0.18	1.99	-33.79	0.12
	$Infit_{new}$	0.99	0.04	0.88	1.09	0.98	0.04	0.83	1.25
	$Infit_{old}$	1.00	0.04	0.90	1.16	1.00	0.06	0.78	1.46
	$Outfit_{new}$	1.01	0.06	0.85	1.34	1.02	0.41	0.81	7.99
	$Outfit_{old}$	1.03	0.07	0.89	1.53	1.23	2.40	0.83	41.78
	r_{pm_new}	0.31	0.06	0.17	0.48	0.31	0.07	0.06	0.53
	r_{pm_old}	0.31	0.06	0.17	0.48	0.31	0.07	0.06	0.53

Table 2.3 presents the number and percentage of misfit and good-fit items in the two samples. Both samples flagged the same items for misfit[2], so only the results from the AllYr sample are presented in this report. As shown in the table, a total of seven and five Reading items (2.46% and 1.75%) and 23 and 20 Mathematics items (7.54% and 6.56%) were flagged for misfit based on the point measure correlations and either the strong or weak criteria, respectively. In other words, at least 97.5% of Reading items and 92.5% of Mathematics items passed the fit check. Items flagged for misfit will be reviewed for content and psychometrics prior to deciding whether to deactivate them.

Table 2.3. Number and Percentage of Misfit and Good-fit Items

Sample	Criteria	Misfit				Good Fit			
		Reading		Mathematics		Reading		Mathematics	
		#Items	%	#Items	%	#Items	%	#Items	%
AllYr	r_{pm}<.2	5	1.75	17	5.57	–	–	–	–
	Infit/outfit>1.2 | Infit/outfit<.8 (Strong)	2	0.70	6	1.97	–	–	–	–
	Infit/outfit>1.5 | Infit/outfit<.5 (Weak)	0	0.00	3	0.98	–	–	–	–
	Good Item Fit (Strong)	–	–	–	–	278	97.54	282	92.46
	Good Item Fit (Weak)	–	–	–	–	280	98.25	285	93.44
	Total #Items Flagged (r_{pm}<.2 + Strong)	7	2.46	23	7.54	–	–	–	–
	Total #Items Flagged (r_{pm}<.2 + Weak)	5	1.75	20	6.56	–	–	–	–

Overall, these results indicate that the fit of the items evaluated in this study were improved with their adjusted RITs based on the 2020 scale alignment study. This finding was the same as that in the 2019 study.

[2] One Mathematics item and one Reading item flagged by AllYr samples were not included in the analysis using the LatestYr sample due to the sample size requirement.

3. References

He, W. (2019). *MAP Growth K–2 item fit analysis study*. NWEA.

Thum, Y., & Kuhfeld, M. (2019). *MAP Growth K–2 to MAP Growth 2–5 temporary re-score solution*. NWEA

Thum, Y., & Kuhfeld, M. (2020). *MAP Growth K–2 item difficulty adjustment: Part 2*. NWEA.