MAP Growth K–2 Item Fit Analysis Study

May 2019

Wei He, Ph.D., NWEA Psychometric Solutions

nwea

Suggested citation: He, W. (2019). *MAP Growth K–2 item fit analysis 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 a recent scale alignment study conducted by NWEA® to realign the scales underlying the MAP Growth K–2 and MAP Growth 2–5 Reading and Mathematics tests (Thum & Kuhfeld, 2019). Part of that study involved adjusting the difficulties of MAP Growth K–2 items (i.e., each item's Rasch Unit (RIT) value) using growth model predictions from longitudinal test results. 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). Specifically, 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.

1.1. Data

Items of interest in this study were MAP Growth K–2 Reading and Mathematics items whose difficulties were adjusted in the scale alignment study. Responses to these items were from the MAP Growth K–2 test events administered in five states between 2010 and 2017. Using the adjusted item difficulties and old item responses, these test events were rescored for this study. Based on the item responses between these years, the item fit analyses were conducted. To investigate the degree to which item fit can be affected by different samples, two samples, AllYr and LatestYr, were constructed:

1. AllYr: Item responses from all the years in which an item was exposed between 2010 and 2017
2. LatestYr: Item responses from the latest year between 2010 and 2017

For example, if an item was used between 2012 and 2015, its AllYr sample included all responses across the four years, but its LatestYr sample included responses from just 2015. Table 1.1 presents the number of items included in this study by year from 2010 to 2017. Items with less than 300 responses were excluded from the study, resulting in a total of 4,680 Reading and 4,650 Mathematics items in the AllYr sample and 4,679 Reading and 4,648 Mathematics items in the LatestYr sample.

Table 1.2 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[1] were slightly smaller than the average old item RITs, with the differences being 3 and 2 RITs for Reading and Mathematics, respectively. For both samples, the correlations between the old and new item RITs were 0.98 for both content areas. As shown in Table 1.3, the average number of responses for Reading and Mathematics items in the AllYr sample are 55,705 and 57,243, respectively, and the average number of responses for Reading and Mathematics items in the LatestYr sample are 9,924 and 10,132, respectively. The AllYr sample has a higher average number of responses because the items in the sample often include responses from multiple years, whereas items in the LatestYr sample only include responses from one year.

[1] The relationship between RIT and logit is RIT=(logit*10)+200 or Logit=(RIT-200)/10.

Table 1.1. Number and Percentage of Items by Year

Year	Reading		Mathematics	
	#Items	%Items	#Items	%Items
2010	19	0.41	16	0.34
2011	1	0.02	29	0.62
2012	223	4.77	261	5.62
2013	89	1.90	73	1.57
2014	205	4.38	166	3.57
2015	61	1.30	268	5.77
2016	419	8.95	531	11.42
2017	3,662	78.26	3,304	71.08
Total	**4,679**	**100.00**	**4,648**	**100.00**

Table 1.2. Descriptive Statistics of Item Difficulties

Content Area	#Items		Item RIT Group	RIT			
	AllYr	LatestYr		Mean	SD	Min.	Max.
Reading	4,680	4,679	New	155	17	114	211
			Old	158	19	110	223
Mathematics	4,650	4,648	New	164	20	111	228
			Old	166	24	107	242

Table 1.3. Descriptive Statistics of Item Responses

Sample	Content Area	#Items	Response Count per Item			
			Mean	SD	Min.	Max.
AllYr	Reading	4,680	55,705	53,853	372	458,953
	Mathematics	4,650	57,243	58,824	305	570,404
LatestYr	Reading	4,679	9,924	8,793	300	90,648
	Mathematics	4,648	10,132	9,607	300	120,515

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 1–3:

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

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

$$r_{pm_i} = \frac{\sum_{n=1}^N (O_{ni} - \bar{O})\left(\widehat{\theta_{ni}} - \bar{\hat{\theta}}\right)}{\sqrt{\sum_{n=1}^N (O_{ni} - \bar{O})^2 \sum_{n=1}^N \left(\widehat{\theta_n} - \bar{\hat{\theta}}\right)^2}} \quad (3)$$

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*.
- $\widehat{\theta_{ni}}$ is the ability estimate of examinee *n* who was administered item *i*.
- $\bar{\hat{\theta}}$ 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 1–3. 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).

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 4 and 5. 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.

$$Infit_{diff_i} = Abs(Infit_{new_i} - 1) - Abs(Infit_{old_i} - 1) \quad (4)$$

$$Outfit_{diff_i} = Abs(Outfit_{new_i} - 1) - Abs(Outfit_{old_i} - 1) \quad (5)$$

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.

Step 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 – Figure 2.4 plot both the new and old infit and outfit statistics against item difficulties for Reading and Mathematics using both the AllYr and LatestYr samples. The new fit statistics, marked in red, were calculated using the adjusted item difficulties, new ability estimates, and item responses. The old fit statistics, marked in green, were calculated using the original item difficulties, ability estimates, and item responses. Three Mathematics items in the AllYr sample with an old outfit value > 6 were excluded in Figure 2.2, and 10 Reading items in the LatestYr sample with an old outfit value > 6 were excluded from
Figure 2.3.

As shown in the figures, the old infit and outfit statistics (green) are more scattered than their new counterparts (red), with more values larger than 2 regardless of the sample. This suggests that the adjustment of item difficulties has helped with the improvement of the item fit. Compared with the Reading items, Mathematics items had more variations regarding the fit statistics. Outfit statistics exhibited more variations than infit statistics regardless of the sample.

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

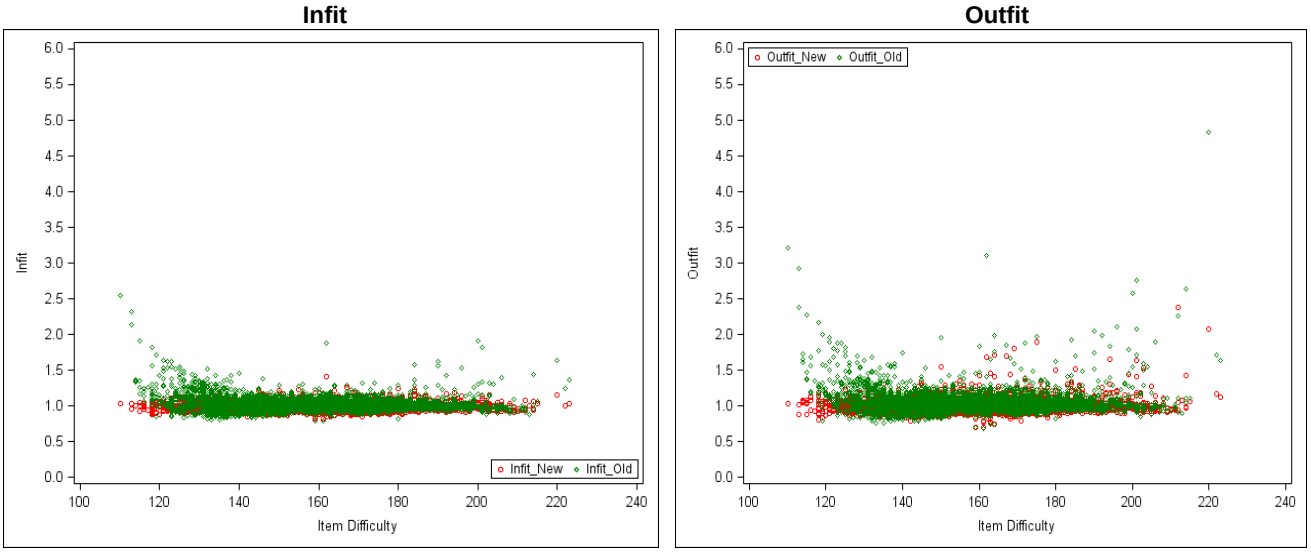

Figure 2.2. New and Old Infit and Outfit Statistics vs. Item Difficulty (AllYr Sample)—Mathematics

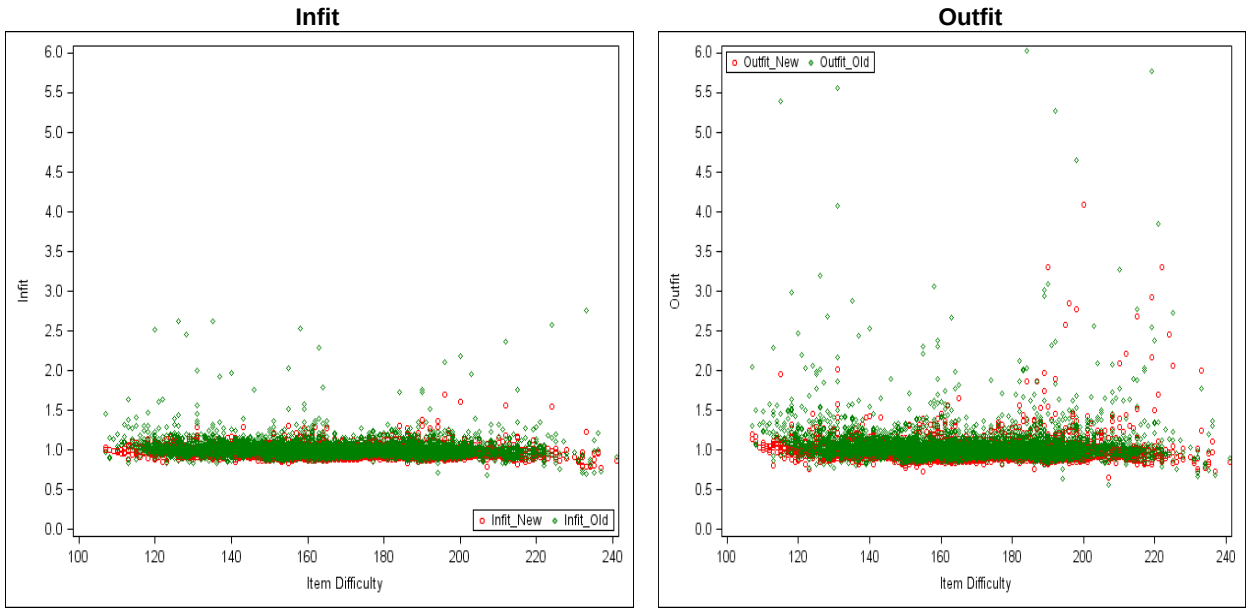

Figure 2.3. New and Old Infit and Outfit Statistics vs. Item Difficulty (LatestYr Sample)—Reading

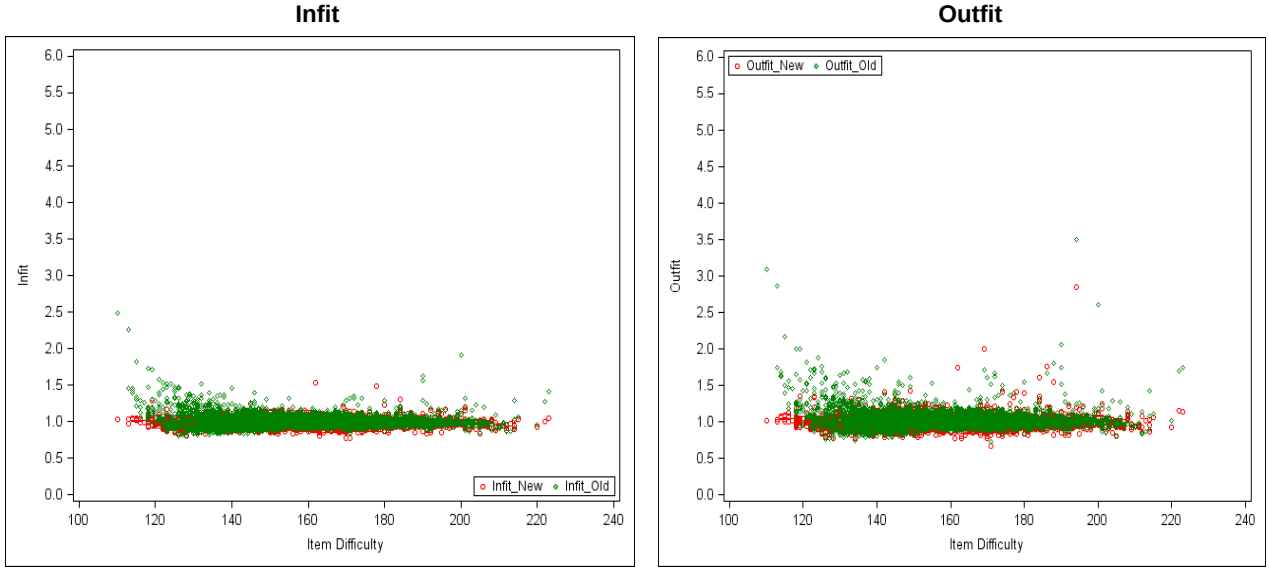

Figure 2.4. New and Old Infit and Outfit Statistics vs. Item Difficulty (LatestYr Sample)—Mathematics

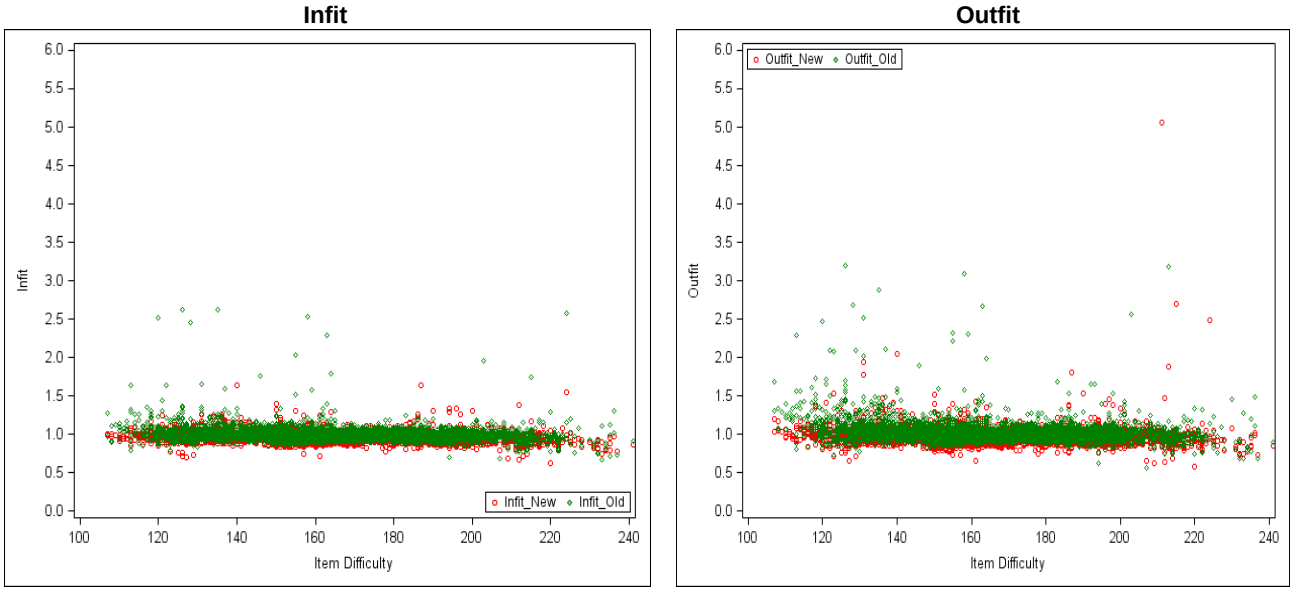

Figure 2.5 – Figure 2.8 plot both the new and old infit and outfit statistics against each other for the AllYr and the LatestYr samples. The new statistics are plotted on the x-axis, and the old statistics are plotted on the y-axis. As shown in the figures for both samples, the old and new statistics are linearly associated but with quite a few outliers, particularly outfit, with old statistics being substantially higher than their corresponding new fit statistics. This indicates the improvement of the item fit by the adjusted item difficulties.

Figure 2.5. New vs. Old Infit and Outfit Statistics (AllYr Sample)—Reading

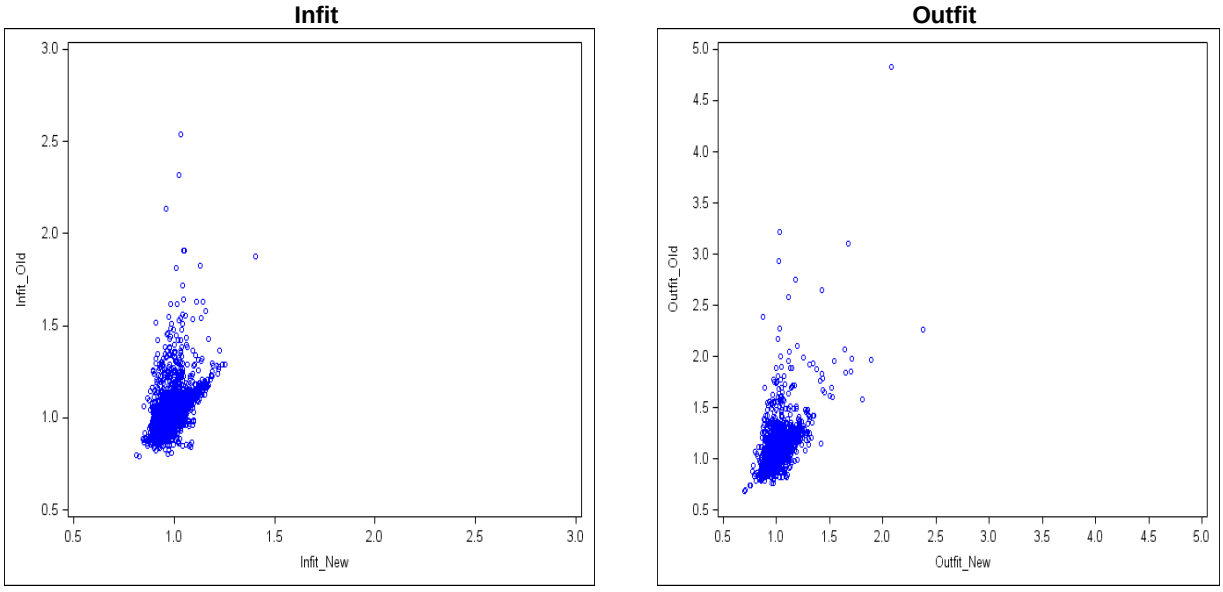

Figure 2.6. New vs. Old Infit and Outfit Statistics (AllYr Sample)—Mathematics

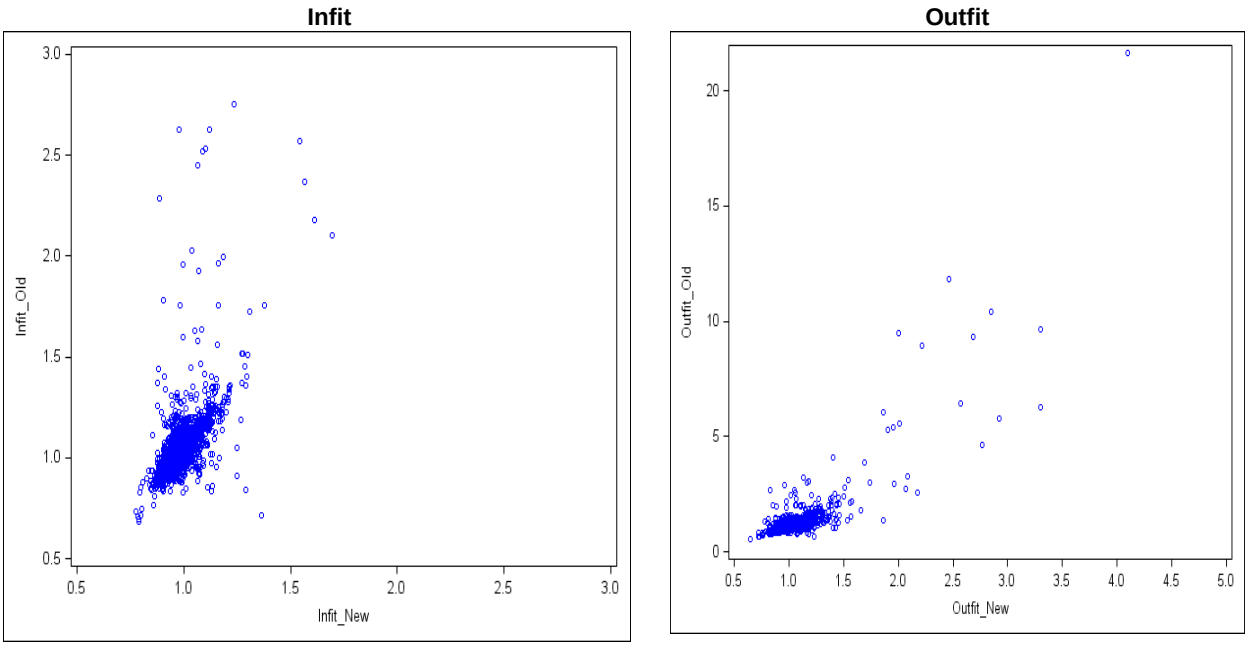

Figure 2.7. New vs. Old Infit and Outfit Statistics (LatestYr Sample)—Reading

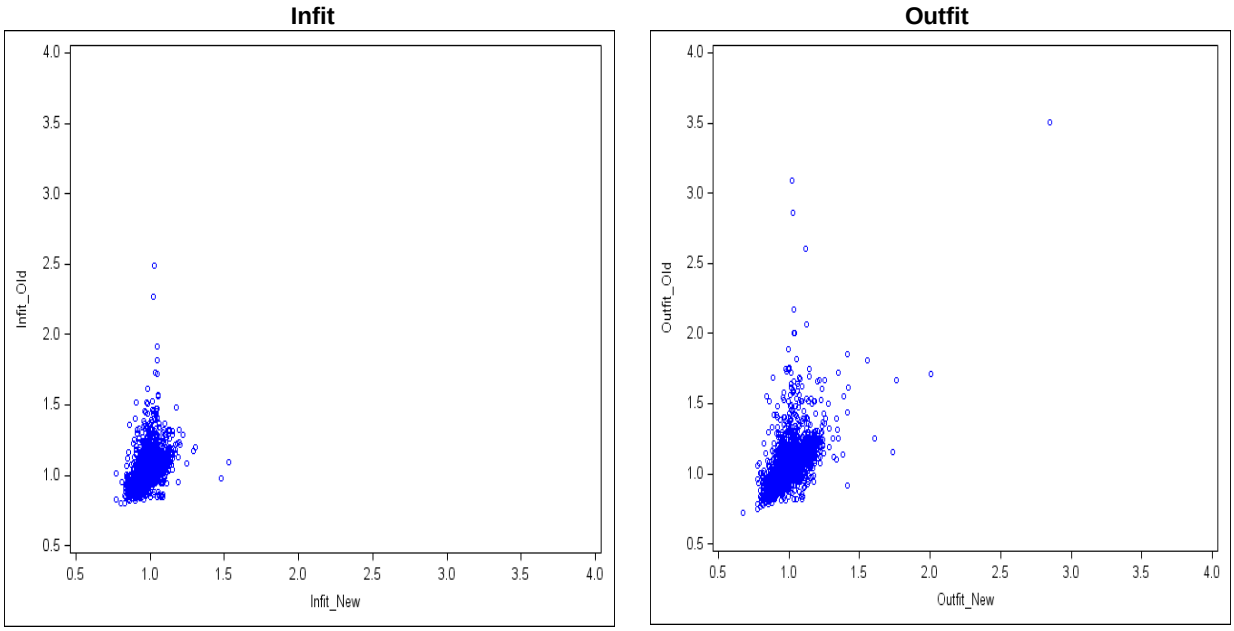

Figure 2.8. New vs. Old Infit and Outfit Statistics (LatestYr Sample)—Mathematics

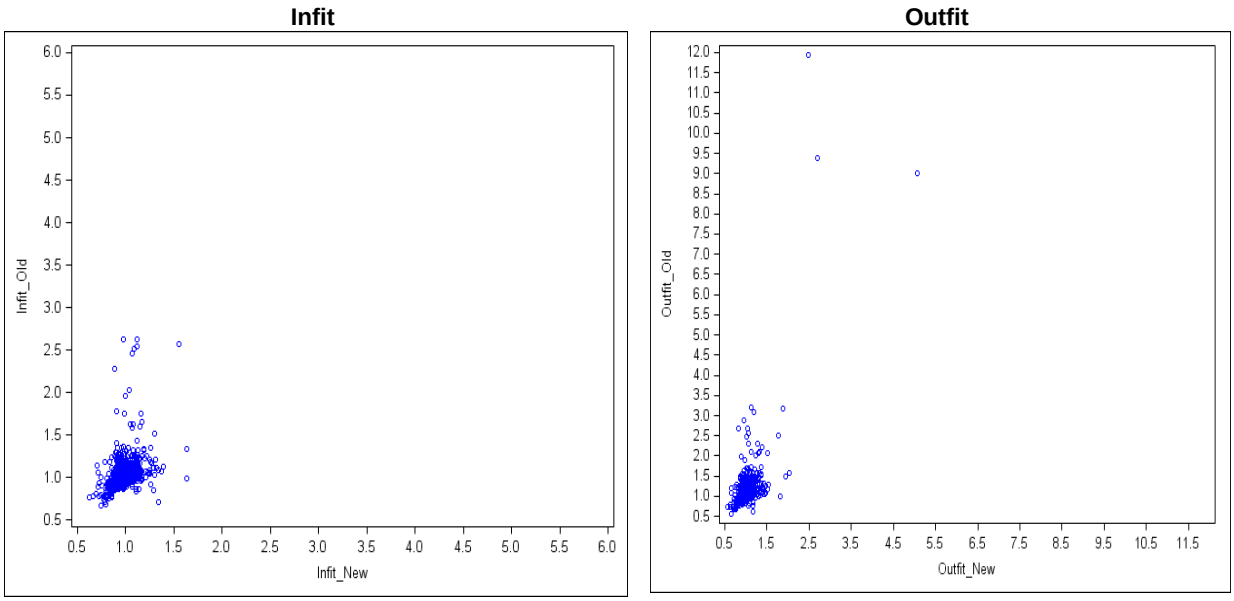

Table 2.1 presents the Pearson correlation coefficients between the new and old infit and outfit statistics for both samples. All coefficients but one are between 0.50 and 0.79, suggesting that the new and old fit statistics are moderately correlated.

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

	r			
	Reading		Mathematics	
Sample	$Infit_{New}$,$Infit_{Old}$	$Outfit_{New}$,$Outfit_{Old}$	$Infit_{New}$,$Infit_{Old}$	$Outfit_{New}$,$Outfit_{Old}$
AllYr	0.53	0.63	0.67	0.81
LatestYr	0.55	0.62	0.54	0.70

Table 2.2 presents the summary infit and outfit statistics for the two samples for the MAP Growth K–2 items, including the point measure correlations, from Step 1 and Step 2 (e.g., mean, standard deviation (SD), and minimum and maximum). 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 results of the Reading items were used to illustrate these findings. For both samples, the absolute differences between the new statistics and 1.00 (i.e., the expected values for both infit and outfit) are smaller than those between the old statistics and 1.00. For the AllYr sample, the average values of the absolute differences between the new infit and outfit statistics and 1.00 are 0.04 and 0.06, respectively, whereas they are 0.06 and 0.09 for the old infit and outfit statistics. For the LatestYr sample, the average values of the absolute differences between the new infit and outfit statistics and 1.00 are 0.04 and 0.05, respectively, whereas they are 0.06 and 0.08 for the old infit and outfit statistics. Overall, these results indicate that the absolute differences between the new statistics and 1.00 are smaller than those between the old statistics and 1.00, which serves as additional evidence that item fit has been improved with the adjusted item difficulties from the scale alignment study.

Table 2.2. Summary Infit and Outfit Statistics

Sample	New and Old Infit and Outfit Statistics	Reading				Mathematics			
		Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
AllYr	$Abs(Infit_{new}-1)$	0.04	0.03	0.00	0.41	0.05	0.04	0.00	0.70
	$Abs(Infit_{old}-1)$	0.06	0.08	0.00	1.54	0.06	0.10	0.00	1.75
	$Abs(Outfit_{new}-1)$	0.06	0.06	0.00	1.38	0.07	0.12	0.00	3.10
	$Abs(Outfit_{old}-1)$	0.09	0.14	0.00	3.83	0.12	0.51	0.00	20.64
	$Infit_{diff}$	-0.02	0.07	-1.51	0.10	-0.01	0.08	-1.60	0.20
	$Outfit_{diff}$	-0.03	0.12	-2.75	0.27	-0.05	0.42	-17.55	0.49
	$Infit_{new}$	0.99	0.05	0.81	1.41	0.97	0.06	0.78	1.70
	$Infit_{old}$	1.02	0.10	0.79	2.54	1.01	0.11	0.68	2.75
	$Outfit_{new}$	1.00	0.08	0.70	2.38	0.99	0.13	0.65	4.10
	$Outfit_{old}$	1.05	0.16	0.68	4.83	1.07	0.52	0.57	21.64
	r_{pm_new}	0.33	0.07	0.06	0.60	0.33	0.07	-0.02	0.55
	r_{pm_old}	0.33	0.07	0.06	0.61	0.33	0.07	-0.02	0.55
LatestYr	$Abs(Infit_{new}-1)$	0.04	0.04	0.00	0.53	0.05	0.05	0.00	0.64
	$Abs(Infit_{old}-1)$	0.05	0.07	0.00	1.49	0.05	0.08	0.00	1.63
	$Abs(Outfit_{new}-1)$	0.06	0.06	0.00	1.85	0.07	0.09	0.00	4.06
	$Abs(Outfit_{old}-1)$	0.08	0.11	0.00	2.51	0.08	0.26	0.00	10.94
	$Infit_{diff}$	-0.01	0.07	-1.46	0.46	0.00	0.08	-1.60	0.63
	$Outfit_{diff}$	-0.02	0.10	-2.06	0.58	-0.01	0.21	-9.46	0.80
	$Infit_{new}$	0.98	0.05	0.77	1.53	0.96	0.06	0.63	1.64
	$Infit_{old}$	1.01	0.09	0.80	2.49	0.99	0.10	0.67	2.63
	$Outfit_{new}$	0.99	0.08	0.67	2.85	0.97	0.11	0.58	5.06
	$Outfit_{old}$	1.03	0.13	0.72	3.51	1.03	0.27	0.57	11.94
	r_{pm_new}	0.32	0.07	0.03	0.61	0.33	0.07	-0.04	0.53
	r_{pm_old}	0.32	0.07	0.03	0.61	0.33	0.07	-0.04	0.54

Table 2.3 presents the number and percentage of misfit and good-fit MAP Growth K–2 items in the two samples. Items flagged based on a low point measure correlation in Step 3 and on strong and weak infit and outfit criteria in Step 4 were exclusive to each other (i.e., items flagged in Step 3 were not included in Step 4), but items flagged based on the weak criteria are a subset of items flagged based on the strong criteria.

Compared with the AllYr sample, the LatestYr sample flagged slightly more items for misfit. For the AllYr sample, a total of 201 and 152 Reading items (4.30% and 3.25%) and 229 and 148 Mathematics items (4.92% and 3.18%) were flagged for misfit based on the point measure correlations and either the strong or weak criteria, respectively. For the LatestYr sample, a total of 214 and 169 Reading items (4.57% and 3.61%) and 280 and 203 Mathematics items (6.03% and 4.37%) were flagged for misfit based on the point measure correlations and either the strong or weak criteria, respectively. This indicates that at least 94% of Reading and Mathematics items passed the fit check. Items with point measure correlations less than 0.2 will be deactivated, and items flagged based on the infit and outfit statistics in each sample 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	146	3.12	128	2.75	–	–	–	–
	Infit/outfit>1.2 | Infit/outfit<.8 (Strong)	55	1.18	101	2.17	–	–	–	–
	Infit/outfit>1.5 | Infit/outfit<.5 (Weak)	6	0.13	20	0.43	–	–	–	–
	Good Item Fit (Strong)	–	–	–	–	4,479	95.70	4,421	95.08
	Good Item Fit (Weak)	–	–	–	–	4,528	96.75	4,502	96.82
	Total #Items Flagged (r_{pm}<.2 + Strong)	201	4.30	229	4.92	–	–	–	–
	Total #Items Flagged (r_{pm}<.2 + Weak)	152	3.25	148	3.18	–	–	–	–
LatestYr	r_{pm}<.2	167	3.57	199	4.28	–	–	–	–
	Infit/outfit>1.2 | Infit/outfit<.8 (Strong)	47	1.00	81	1.74	–	–	–	–
	Infit/outfit>1.5 | Infit/outfit<.5 (Weak)	2	0.04	4	0.09	–	–	–	–
	Good Item Fit (Strong)	–	–	–	–	4,465	95.43	4,368	93.98
	Good Item Fit (Weak)	–	–	–	–	4,510	96.39	4,445	95.63
	Total #Items Flagged (r_{pm}<.2 + Strong)	214	4.57	280	6.02	–	–	–	–
	Total #Items Flagged (r_{pm}<.2 + Weak)	169	3.61	203	4.37	–	–	–	–

To investigate the degree of consistency that both samples flag misfit items, Table 2.4 presents the number of items flagged for misfit by both samples (i.e., the number of common items) and the number of items flagged for misfit in only one of the samples (i.e., the number of unique items) based on the different criteria. If the samples are representative of the population, the same items are expected to be flagged by both samples. However, for both content areas, only about half of the misfit items were flagged as common in both samples. For Reading, 141 out of the 274 unique misfit items were common in both samples, whereas for Mathematics, 159 out of the 350 unique misfit items were common in both samples.

Table 2.4. Number of Misfit Items Flagged Commonly by Both Samples and Uniquely by Each Sample

Content Area	Criteria	#Misfit Items			
			Unique		
		Common	AllYr	LatestYr	Total
Reading	r_{pm}<.2	115	31	52	198
	Infit/outfit>1.2 | Infit/outfit<.8 (Strong)	15	40	32	87
	Infit/outfit>1.5 | Infit/outfit<.5 (Weak)	0	6	2	8
	r_{pm}<.2+Strong	141 (51%)	60	73	274
Mathematics	r_{pm}<.2	112	16	87	215
	Infit/outfit>1.2 | Infit/outfit<.8 (Strong)	30	71	51	152
	Infit/outfit>1.5 | Infit/outfit<.5 (Weak)	1	19	3	23
	r_{pm}<.2+Strong	159 (45%)	70	121	350

3. Summary and Conclusion

This study examined the fit of MAP Growth K–2 items involved in the previous scale alignment study that used the growth modeling approach to predict person abilities and adjusted item difficulties based on the predicted person abilities (Thum & Kuhfeld, 2019). Two different samples were constructed in the study: AllYr and LatestYr. For each sample, a set of fit statistics including infit, outfit, and point measure correlations were calculated for each item based on the old and new values for the item difficulties and the ability estimates. The overall results of the study indicate that, regardless of the samples, a substantial number of items for both content areas (at least 94%) exhibit good model fit after their item difficulties were adjusted by the scale alignment study.

However, the study also found that the items flagged for misfit by both samples were quite different from each other, which was surprising. One possible reason is the change of the curriculum throughout the years. Over the past decade, more and more states have adopted the Common Core State Standards (CCSS), which are more stringent with the emphasis on applying what students learn to solve real-life problems. As a result, some of the items developed before the CCSS are likely to become less relevant to the curriculum and, thus, have poorer fit. Another possible reason is that some of the item content has become obsolete, and such items have become less fit than before. Sample differences could also be a reason. It is likely that the sample in the LatestYr, a subset sample of AllYr, contained students with demographic information different from that in the AllYr. Unfortunately, the limitations of the data (e.g., no demographic information) prevented any actions to be taken to explore this further. Regardless, NWEA plans to closely monitor these items and take necessary actions such as examining item fit on a regular basis to ensure that only items exhibiting adequate data fit are used in the MAP Growth K–2 tests.

4. References

Thum, Y. M. & Kuhfeld, M. (2019). *Realigning the scale of one item pool to another: IRT linking using growth data*. NWEA.