Content Proximity Spring 2022 Pilot Study Research Brief

J. Patrick Meyer, Ann Hu, Sylvia Li
Psychometric Solutions
June 2023

nwea

Document History

Date	Version	Description
2023-05-24	0.1	Initial draft created
2023-06-23	1.0	Finalized by Patrick Meyer; published

Table of Contents

1. Introduction

The Content Proximity Project was designed to improve the content validity of the MAP® Growth™ assessments while retaining the ability for the test to adapt off-grade and meet students wherever they are in their learning. Two main features of the project were the development of an enhanced item selection algorithm, and a spring pilot study conducted in volunteer school districts. The purpose of the pilot study was to evaluate the new algorithm during live testing, study the comparability of scores with traditional MAP Growth assessments, and produce evidence of test content validity and score reliability. The pilot study began in spring 2022 with a group of NWEA Partners who volunteered to participate.

The Content Proximity Project was initiated with several benefits in mind. The primary benefits are enhanced content validity, improved perceptions of test quality, and greater test taking engagement. The test will continue to adapt off grade when needed to deliver items of suitable difficulty for a student. However, this adaptation will be done in such a way that test events will be more closely aligned with grade-level content, especially for students exhibiting typical performance for a grade. The stronger preference for grade-level content means that the test more closely matches the subject matter students have an opportunity to learn in school. Subsequently, MAP Growth scores should allow for better connections to curriculum materials and resources, and produce scores that are more highly correlated with end-of-year summative tests.

1. Item Selection Algorithm

The Content Proximity Project aimed to improve content validity by delivering test events that satisfy updated test blueprints, and increases the preference for on-grade items. This goal required an enhancement to the MAP Growth item selection algorithm.

Test design aims to satisfy statistical and content requirements. Balancing these two aspects of a test is straightforward in a linear test such that a human test designer chooses items for a test and the test form is the same for everyone. The test does not change or adapt for each student. Test design is more complicated in CAT because each test must meet statistical and content requirements while being tailored to an individual examinee. A sequential item selection algorithm dynamically creates a test by choosing items one at a time, according to an examinee’s responses to previously administered items. The algorithm must decide on-the-fly how to select an item for a test and ultimately satisfy the intended statistical and content requirements. Psychometricians have developed a variety of item selection algorithms for balancing statistical and content requirements, and each one has strengths and weaknesses.

Statistical and content requirements often work against each other. Improving the statistical aspects of a CAT may reduce the ability of a test to satisfy content requirements, and vice versa. It is a manifestation of the long-standing reliability/validity tradeoff in measurement where test design features that improve reliability can limit the type of inferences about student ability and evidence supporting those inferences.

1.1 Statistical Requirements

Statistical requirements are derived from the underlying item response model and address the measurement precision and reliability aspects of a CAT. Maximum information item selection is a procedure where the most informative item is selected at each iteration. The ultimate result is a test that yields maximum information about an examinee's ability (test score) for the given number of items.

MAP Growth uses the Rasch item response model for test scaling and calibration. For the Rasch model, an item provides maximum information about the examinee's ability when the probability of a correct response is 0.5. This probability occurs when item difficulty and examinee ability are equal (i.e., $\theta = \delta$).

Item information is summed over all items on a test to compute the total test information at a given ability level. Test information is inversely related to the standard error of measurement (SEM) for an examinee's ability. The SEM decreases as test information increases. The larger the test information, the greater the precision of an examinee's ability estimate. For a group of examinees, score reliability is a function of the variance of their ability estimates, and the average of their squared SEMs. For the same amount of ability variance, reliability increases as SEM decreases for examinees. Thus, selecting items that maximize information for each examinee ultimately leads to high levels of score reliability for a group of examinees.

The goal of maximum information item selection is to create a test with the largest possible information (i.e. lowest possible SEM) for a given number of items. However, the maximum information approach to item selection only accounts for a test's measurement precision. It does not incorporate any aspect of test content outside of its relationship to item difficulty. Item selection must be redefined to explicitly account for test content requirements.

1.2 Content Requirements

Content requirements are defined by test blueprints. They are not part of an item response model and cannot be controlled through an algorithm that solely relies on maximum information item selection. Maximum information item selection must be augmented in some way, or the optimization problem must be redefined altogether to produce a test event that satisfies content requirements.

1.3 MAP Growth Item Selection Algorithms

1.3.1 Constrained CAT

Constrained CAT (C-CAT; Kingsbury & Zara, 1989) is an algorithm that works by partitioning the item pool into mutually exclusive content categories, identifying a category that has not reached its target number of items, and then selecting the most informative item from the category. Randomesque exposure control can be added by randomly selecting from a group of the most informative items (Kingsbury & Zara, 1989) in the category. MAP Growth tests have used C-CAT, since the assessments were offered in computer adaptive format. C-CAT is effective with a few mutually exclusive content categories. When items have multiple and overlapping content assignments, partitioning the item pool into groups may result in empty or very sparse partitions resulting in over exposure or item starvation.

1.3.2 Algorithm for Reward Listed Item Selection (ARLIS)

The enhanced item selection algorithm was named the Algorithm for Reward Listed Item Selection (ARLIS). It is a compensatory version of the maximum priority index (Cheng & Chang, 2009) that borrows ideas from reinforcement learning. It may be thought of as reinforcement learning with a greedy policy and frequent deterministic rewards. In reinforcement learning, an *agent* chooses an *action* based on the current state of the *environment* to earn a *reward*. The agent only knows the current state and the available actions. The agent's goal is to maximize the total reward. In the context of CAT, an item selection engine is the agent that chooses an action of selecting a particular item from the pool. The environment is everything that is not the agent. It includes the items available in the pool, the items already selected for the test, the examinee, and other aspects of the test administration. The environment is constantly changing, and the agent must know how to choose the right actions.

The approach we have taken with ARLIS is to assign a reward to each content feature. After calculating the reward for each individual feature, a total reward is computed for each item. The total reward is a weighted average of functions representing statistical and content requirements. The total reward is calculated for each item in the pool. Items are then listed in descending order of total reward (i.e., reward listed), and the item with the largest value is selected. If multiple items tie for the largest value, then one is randomly selected from them. Having tied values is ideal from the standpoint of exposure control because the random selection of an item results in a randomesque exposure control mechanism (Kingsbury & Zara, 1989).

1.4 Grade/Difficulty Tradeoff Research Study

A primary question about ARLIS is how it adapts off grade. This aspect can be addressed analytically by considering reward functions for item information and item grade (details not shown). Specific analytic results depend on the configuration of the item grade reward. For one configuration, an on-grade item with a difficulty 11 RIT different from the student's ability is as likely to be selected as a maximally informative off-grade item. As the absolute difference between student ability and item difficulty increases beyond this point, the off-grade item becomes preferred and has a higher total reward. Furthermore, a much larger difference between student ability and item difficulty is needed for an item that is two grades below the student's grade to be as preferred as a maximally informative off-grade item.

In practice, the total reward will include functions for other content features, and it will not simply involve item grade and item difficulty. The array of content features may cause the algorithm to select off-grade items more frequently if no on-grade items with the same difficulty and content features exist. Likewise, the algorithm may continue to prefer an on-grade item that has a preferred content feature even though an off-grade item may be more informative. A key factor that affected the degree of off-grade adaption was the difference between the group mean and the item pool mean for an instructional area. The greater the difference, the more frequently off-grade items were selected.

2. Content Proximity Pilot Study

The Content Proximity Pilot Study began in spring 2022 with three school districts that agreed to participate in the study. Two school districts were existing MAP Growth users with prior MAP Growth data from the Common Core State Standards (CCSS) aligned test. The other district was a new partner with no prior MAP Growth data who was interested in the CCSS aligned test. As explained below, the new district was excluded from the analysis because of not having prior MAP Growth data.

Analysis of data from the pilot study addressed several aspects of the assessment and test scores. It included an evaluation of test content and the comparability of test scores, person and item fit statistics, and norms predictions.

2.1 Pilot Study Sample and Propensity Score Matching

The pilot study focused on grades K-8. Comparison schools taking existing MAP Growth tests aligned to the CCSS were selected through propensity score matching, given the convenient nature of the sample. Propensity score matching (Rosenbaum & Rubin 1983, 1985) is a statistical method that uses covariates to predict the probability of receiving the treatment. It is a way of adjusting for bias in the outcome due to nonrandom selection of participants and making the treatment and comparison groups comparable on covariates observed prior to intervention. In the present case, the “treatment” was defined as participation in the pilot study group and taking a CCSS Content Proximity test, and “nontreatment” was defined as taking an existing CCSS MAP Growth test.

The propensity score analysis was conducted separately for each grade and subject. Matching variables included student sex, race, winter RIT score, response time effort, and school challenge index. The pilot and comparison groups each involved about 1,000 student per grade across grades K-8. Demographics for each group were nearly identical because of the matching process.

2.2 Test Event Content Representation

2.2.1 Alignment of Items to Student Grade

The Content Proximity project is aimed at increasing the content validity of MAP Growth tests by emphasizing on-grade content and proportionally representing content entailed by a set of curriculum standards. While grade-level instruction is suitable for most students, not every student enters a grade with the same amount of prior achievement. Some students may need more time learning prerequisites for grade-level standards. Still other students may be ready for learning material that is typical of standards at higher grade levels. MAP Growth now gives greater preference for on-grade content, but this preference is balanced with the need to deliver off-grade content when appropriate for a student. Thus, the three key claims for enhanced content validity are that (a) test events for typical students will have a large majority of items representing on-grade content, (b) items with a grade level closer to a student’s grade will be more common than items with a more distal grade, and (c) test events for very low- and high-performing students will have more items representing off-grade content; the off-grade items will be below-grade items for low achieving students, and above-grade items for high achieving students.

Although details are not provided in this research brief, results from the pilot study showed that:

1. There were significantly more on-grade items in the pilot study group. The percentage of on-grade items in math increased by at least 20% in grade 3-8. The percentage of on-grade items in reading increased by at least 20% in all grades.

2. The enhanced item selection algorithm makes better use of adjacent-grade items. Items close to the student grade are more common than those further away. That is, it selects items from one grade away from the student’s grade more often than it selects items two

grades away. In the comparison group, similar percentages of off-grade items are seen for items one, two, or even three grades away.

Table 2.1 provides more information about on- and off-grade items for each student grade level. It shows the number of item grades represented by 5% or more items for each student grade. The analysis was done for all students in each grade, and also by dividing students into three ability groups according to their RIT score on the spring test. The lower 10% are students in the first decile, and the upper 10% are students in the 10th decile. The middle 80% are students between the first and tenth decile.

As shown in the table, the number of grades represented by items in the pilot study were substantially less than the number of grades represented in the comparison group. For example, only two grade levels were represented by items in grade 3 math pilot group, but 6 grade levels were represented by items in the comparison group (relevant numbers marked in Table 2.1 with superscript 1). Test content for the pilot group largely represented on- or adjacent-grade content. As another example, the middle 80% of students in the grade 3 Pilot group had items that spanned three grade levels, whereas the middle 80% of grade 3 student in the Comparison group had items that spanned six grade levels (relevant numbers marked with superscript 2).

As noted in the content validity claims, there should be more off-grade items (i.e. more grade levels represented) for low- and high-performing students and fewer grade levels represented for “typical” students, if the test is adapting off-grade for low- and high-performing students. Information in the table support this trend for the pilot study group. It indicates that more grades were represented by items for in the lower and upper pilot study groups than for the middle 80%. In the comparison group, the middle 80% tended to have more grades represented than either the lower or upper 10%. That is, the Comparison group showed the opposite of what was expected for the content validity claims.

Table 2.1 *Number of item grades represented with 5% or more items*

		Pilot				Comparison			
Subject	Student Grade	Lower 10%	Middle 80%	Upper 10%	All Students	Lower 10%	Middle 80%	Upper 10%	All Students
Math	1	3	2	3	2	3	4	3	3
	2	2	2	2	3	4	3	2	3
	3	2	2	2	2[1]	4	6	5	6[1]
	4	2	1	2	2	5	6	5	5
	5	4	3	2	3	5	5	5	5
Reading	1	2	1	2	1	2	4	3	3
	2	2	2	3	3	3	3	3	3
	3	3	3[2]	4	3	4	6[2]	5	7
	4	3	1	3	3	4	6	5	6
	5	4	3	4	3	5	7	5	6

In math and reading, results for the pilot group showed stronger evidence for the content validity claims related to grade level alignment. The expected pattern of on- and off-grade items was more predominant for the pilot group.

2.2.2 Alignment of Items to Test Blueprint Targets
Test blueprints indicate the number of target items per instructional area in addition to other content requirements. The blueprints for the pilot and comparison groups are different. Therefore, this part of the content analysis focuses exclusively on test events for the pilot group.

Analysis of extant test events from the Content Proximity project showed that test events for the pilot group largely fulfilled blueprint requirements. In the K-2 band for math, over 95% of test events achieved the target number items per instructional area and 100% were within 1 item of the target number. In the 2-5 and 6+ grade bands, at least 50% were within 1 item of the target number, and 90% or more test events were within 2 items of the target number per instructional area.

For the K-2 Reading test, 82% to 84% of test events were within one item of the target number for each instructional area, and 98% were within two items. The number of items per instructional area was never more than three items away from the target. Reading tests for the 2-5 and 6+ grade bands showed more variability in meeting target counts per instructional area. The percentage of test events within two items of the target number for each instructional area ranged from 65 to 80. The percentage within three items of the target ranged from 81 to 93. Fewer test events meeting target item counts per instructional area may be due to the use of more item sets with reading passages. After an item set was selected, items within the set might not cover all instructional areas of interest. More items would be selected from the available instructional areas regardless of whether they were needed or not. As noted below, test blueprints for the 2-5 and 6+ tests intentionally included more reading passages containing item sets.

A goal of the new item selection algorithm was to have more test events with two reading passages (i.e., item sets) on the 2-5 test. Each passage would now be required to have four items per passage. In addition, the target number of reading passages was increased from two to three on the 6+ tests. The enhanced algorithm was successful in meeting these goals. A total of 90% to 97% of pilot group test events on the 2-5 test had two reading passages. Only 6% of events had no reading passages. In the comparison group, the percentages of test events with two reading passages were lower and ranged from 71 to 91. Moreover, 19% of test events in the comparison group had no reading passages.

For the 6+ test, the percentage of pilot group test events with three reading passages ranged from 95-97, and 9% of test events included four reading passages. In the comparison group, no test event had more than two passages. However, test blueprints for the comparison group only specified up to two reading passages. The comparison group still had 12% of test events with no reading passages. The results indicate that the enhanced algorithm was successful at delivering more reading passages per test events as indicated by the updated blueprints.

2.3 Score Comparability Results
The score comparability analysis focused on reliability estimates, descriptive statistics, and multiple regression. The analysis was done separately for each grade. The regression model

used a student's spring test score as the dependent variable. The other independent variables were the same ones used as independent variables in the propensity score matching analysis. They were included in the regression model to refine the similarity among the groups. The variables were a series of demographic variables, response time effort scores, winter RIT scores (i.e., prior achievement) and the school challenge index. All variables and their regression coefficients are listed in Appendix A. The main independent variable of interest was a dummy coded indicator student group (pilot or comparison).

2.3.1 Math scores

Table 2.2 shows test score reliability estimates and descriptive statistics for the math assessments. All reliability estimates for both groups were above 0.9. Reliability estimates ranged from 0.93 in kindergarten to 0.97 in several other grades for the pilot group in spring. These results were very comparable to reliability estimates for the comparison group which ranged from 0.94 to 0.97. Interestingly, the math assessments were shorter for the pilot group than the comparison group, yet reliability estimates were very similar.

Descriptive statistics for math in Table 2.2 show that the two groups have similar average score and standard deviations in winter when both groups took the same assessment. Standard deviations remained similar in the spring test when the pilot group took the Content Proximity tests. However, the average scores for the pilot group in each grade were consistently higher than those for the comparison group. The size of the difference tended to increase as grade increased.

Regression analysis for math indicated that the increased math scores for the pilot group are statistically significant at the 0.05 level for each grade (see Table 2.4). The difference ranged from a low of 0.76 adjusted RIT in grade 2 to a high of 6.42 adjusted RIT in grade 6. As suggested by the descriptive statistics the difference tended to be larger for the 6+ MAP Growth assessments than either the K-2 or 2-5 assessments. Hedge's g estimates of effect size shown in Table 2.4 range from 0.05 in Grade 2 to 0.37 in Grade 6. These values indicated nil or negligible effects to small effects, according to Cohen's (1988) guidance on interpreting effect sizes. In the context of the RIT scale, which has a standard deviation of 10, the larger effect sizes in Table 2.4 indicate a difference of about 2 to 3 RIT points.

2.3.2 Reading scores

Reliability estimates for reading assessments were also nearly identical for both groups (see Table 2.3). Estimates ranged from 0.93 to 0.97 across all grades for both groups. Test length did not change for reading, so reliability estimates were expected to be similar.

Reading scores for both groups in the spring were similar on average. In some cases, the mean reading scores for the Pilot group were slightly larger than those for the comparison group. In other cases, the comparison group means were slightly larger. While the differences in means were very small, the standard deviations were slightly smaller for the Pilot group than they were for the comparison group, particularly for the grades covered by the MAP Growth 6+ assessment.

The regression analysis indicated that the pilot group scored higher than the comparison group for some grades, but lower in others (see Table 2.4Table 2.4). There was no consistent trend in the differences of mean scores in reading scores like there was for the math. In addition, the coefficient was statistically significant at the 0.05 level in only four of the nine regression analysis. One of the significant differences favored the comparison group while the other three favored the pilot group. Effects sizes in Table 2.4 show that the differences were very small and negligible. The largest effect size was 0.07, which was well below the threshold of 0.2 for a small effect. It indicates that the largest mean difference was less than a single RIT point.

Table 2.2. Descriptive statistics for Pilot and Comparison groups taking the MAP Growth math assessment

		Pilot				Comparison			
Term	Grade	N	Reliability	Mean	S.D.	N	Reliability	Mean	S.D.
Winter	K	1,072	0.92	151.56	11.15	1,071	0.94	151.55	11.76
	1	1,034	0.93	166.43	11.94	1,034	0.95	166.59	13.33
	2	1,134	0.95	181.61	13.29	1,134	0.95	181.49	13.55
	3	1,097	0.95	192.35	13.37	1,098	0.95	192.57	13.90
	4	1,186	0.95	199.85	13.41	1,185	0.96	199.63	14.35
	5	1,219	0.95	209.41	14.18	1,220	0.96	210.60	15.72
	6	939	0.95	214.46	14.53	938	0.96	214.52	15.05
	7	903	0.96	220.56	15.15	904	0.96	220.26	15.61
	8	1,001	0.96	225.40	15.75	1,001	0.97	225.46	17.88
Spring	K	1,071	0.93	160.74	11.46	1,071	0.94	159.59	12.20
	1	1,035	0.95	175.70	13.38	1,035	0.95	174.87	13.77
	2	1,133	0.95	190.75	14.35	1,133	0.96	189.96	15.28
	3	1,098	0.96	202.21	15.34	1,098	0.96	198.52	14.63
	4	1,185	0.96	209.40	16.06	1,185	0.97	205.37	16.00
	5	1,220	0.97	217.19	16.90	1,220	0.97	214.90	17.89
	6	938	0.97	224.47	16.89	938	0.97	218.31	16.26
	7	904	0.96	226.81	16.49	904	0.97	223.46	16.90
	8	1,001	0.97	232.74	18.31	1,001	0.97	227.21	18.49

Table 2.3. Descriptive statistics for Pilot and Comparison groups taking the MAP Growth reading assessment

		Pilot				Comparison			
Term	Grade	N	Reliability	Mean	S.D.	N	Reliability	Mean	S.D.
Winter	K	1,067	0.91	147.11	10.33	1,066	0.91	146.95	10.63
	1	1,036	0.94	161.35	13.19	1,038	0.95	160.73	13.86
	2	1,115	0.94	175.07	13.83	1,114	0.96	175.58	15.41
	3	1,101	0.97	188.41	18.13	1,102	0.97	187.85	17.91
	4	1,184	0.96	197.23	16.81	1,183	0.96	197.31	16.65
	5	1,218	0.95	205.86	14.88	1,219	0.96	205.94	15.43
	6	913	0.95	211.74	14.19	912	0.95	211.14	15.07
	7	891	0.95	217.05	13.99	892	0.95	217.39	14.34
	8	971	0.94	221.16	13.52	971	0.95	221.79	14.60
Spring	K	1,066	0.93	155.05	12.02	1,066	0.93	154.50	11.97
	1	1,037	0.95	168.16	13.96	1,038	0.95	168.01	14.8
	2	1,114	0.94	179.87	13.73	1,114	0.95	180.95	14.95
	3	1,102	0.97	193.38	17.89	1,102	0.97	192.10	17.64
	4	1,183	0.96	200.25	16.40	1,183	0.96	199.84	17.18
	5	1,219	0.95	207.97	14.93	1,219	0.96	207.20	16.11
	6	912	0.95	212.48	13.95	912	0.96	212.37	15.47
	7	892	0.95	218.14	14.04	892	0.95	217.67	15.14
	8	971	0.94	221.93	12.98	971	0.95	220.93	15.30

Note: Pilot and comparison samples may differ because students with multiple test events were removed. Students had multiple test events when they switched schools and were tested twice in the same term.

Table 2.4. Summary of main pilot study effects using matched comparison group

Subject	Grade	Estimate, $\hat{\beta}$	p-value	Effect Size
Math	K	1.134	< 0.001	0.098
	1	1.024	0.002	0.061
	2	0.763	0.012	0.053
	3	3.883	0.000	0.246
	4	3.772	0.000	0.251
	5	3.574	0.000	0.132
	6	6.415	0.000	0.372
	7	3.065	0.000	0.201
	8	5.660	0.000	0.301
Reading	K	0.442	0.203	0.046
	1	-0.365	0.286	0.008
	2	-0.600	0.035	-0.075
	3	0.777	0.015	0.072
	4	0.414	0.212	0.024
	5	0.761	0.009	0.050
	6	-0.415	0.217	0.007
	7	0.662	0.060	0.033
	8	1.594	0.000	0.070

2.4 Growth Norms Predicted to Actual RIT Score Comparison

A within-student analysis was conducted by comparing each student’s observed RIT score for spring to the score predicted from prior test scores by the MAP Growth norms model (Thum & Kuhfeld, 2020). The root mean squared difference (RMSD) was calculated as the root mean squared difference between the observed and predicted scores. Results in Table 2.5 showed that for math, the RMSDs for the pilot group were consistently larger than the RMSDs for the comparison group. This result was consistent with the difference in actual math scores observed for the students in the pilot study. The correlations between the actual and predicted scores were high and similar across the two groups. However, the RMSD indicated a slight shift in math scores.

In reading, the predicted and observed scores for the pilot group were similar to the comparison group. In some cases, the RMSDs were lower and the correlations were higher for the pilot group than the comparison group. Predictions of reading scores were similar for both groups.

Table 2.5. Growth norms predicted and observed score comparison

		Pilot			Comparison		
Test	Grade	N	Cor.	RMSD	N	Cor.	RMSD
Math	K	1,071	0.82	11.02	1,071	0.81	10.55
	1	1,035	0.81	11.76	1,035	0.84	10.80
	2	1,133	0.88	11.49	1,133	0.88	10.93
	3	1,098	0.88	11.91	1,098	0.90	8.54
	4	1,185	0.87	12.08	1,185	0.90	8.88
	5	1,220	0.88	10.96	1,220	0.92	8.21
	6	938	0.88	12.67	938	0.92	7.34
	7	904	0.89	9.64	904	0.92	7.14
	8	1,001	0.88	11.24	1,001	0.93	6.84
Reading	K	1,066	0.80	10.41	1,066	0.73	10.86
	1	1,037	0.82	10.24	1,037	0.85	10.33
	2	1,114	0.87	8.24	1,114	0.90	8.36
	3	1,102	0.90	9.14	1,102	0.89	9.04
	4	1,183	0.87	8.70	1,183	0.87	8.73
	5	1,219	0.86	7.86	1,219	0.88	7.81
	6	912	0.86	7.36	912	0.87	7.80
	7	892	0.83	8.03	892	0.84	8.18
	8	971	0.84	7.46	971	0.85	8.19

2.5 Person and Item Fit

Thum and Kingsbury (2017) described three criteria for evaluating the quality of an equal interval vertical scale. They focused on *predictions* concerning:

1. overall scale performance within and across grades,
2. student performance within and across grades, and
3. item performance within and across grades.

Fit statistics provide a way to evaluate these criteria. Two types of fit are the focus of this analysis: (a) the similarity of observed and expected proportion, and (b) the outfit statistic. Both types are calculated for persons (i.e., students) and items.

Observed and expected proportions will be the same, and the difference between observed and expected proportions will be zero when the Rasch model fits the data. The degree of misfit increases as the differences increases in absolute value. The outfit statistic is a common fit statistic in the Rasch measurement literature. It has an expected value of 1. Values larger than one indicate misfit with larger values indicating more misfit. Values less than one are indicative of overfitting (i.e., data fit too well), but are not problematic for measurement.

2.5.1 Person fit

Table 2.6 shows the mean difference between the observed and expected proportion correct scores for the pilot and comparison groups. The mean values are very similar for each group and all are close to zero. The standard deviations are also quite similar. Although not shown here, the difference in observed and expected scores for each RIT score decile was plotted.

The distributions in each decile for the pilot group were similar to the comparison group. Both groups tended to show larger difference in observed and expected proportions correct in the first and last deciles.

Table 2.6. Summary of the average difference between observed and expected proportion correct scores for examinees

		Pilot			Comparison		
Term	Grade	N People	Mean	S.D.	N	Mean	S.D.
Math	K	1,071	0.00	0.01	1,071	0.00	0.01
	1	1,035	0.00	0.01	1,035	0.00	0.01
	2	1,133	0.00	0.01	1,133	0.00	0.01
	3	1,098	0.00	0.01	1,098	0.00	0.01
	4	1,185	0.00	0.01	1,185	0.00	0.01
	5	1,220	0.00	0.01	1,220	0.00	0.01
	6	938	0.00	0.01	938	0.00	0.01
	7	904	0.00	0.01	904	0.00	0.01
	8	1,001	0.00	0.01	1,001	0.00	0.01
Reading	K	1,066	0.00	0.01	1,066	0.00	0.01
	1	1,037	0.00	0.01	1,037	0.00	0.01
	2	1,114	0.00	0.01	1,114	0.00	0.01
	3	1,102	0.00	0.02	1,102	0.00	0.02
	4	1,183	0.01	0.02	1,183	0.01	0.02
	5	1,219	0.00	0.02	1,219	0.00	0.02
	6	912	0.01	0.02	912	0.01	0.02
	7	892	0.01	0.02	892	0.01	0.02
	8	971	0.01	0.02	971	0.00	0.02

Person outfit statistics in Table 2.7 show that the average value close to the expected value of one. The mean outfit statistics are very similar among the two groups in both subjects. The standard deviation is larger for the pilot group in math. The outfit statistic is known to be sensitive to outliers. It may be that outliers are affecting the outfit standard deviation.

Person outfit statistics were also plotted for each score decile, but the results did not reveal any clear difference among the statistics for the two group. As such, the plots are not shown herein.

Table 2.7. Summary of person outfit statistics

Term	Grade	Pilot			Comparison		
		N People	Mean	S.D.	N	Mean	S.D.
Math	K	1,071	0.96	0.15	1,071	0.97	0.19
	1	1,035	1.02	0.51	1,035	0.97	0.24
	2	1,133	1.06	1.19	1,133	0.98	0.17
	3	1,098	0.99	0.14	1,098	0.99	0.12
	4	1,185	1.03	0.27	1,185	1.00	0.13
	5	1,220	1.01	0.20	1,220	1.01	0.14
	6	938	1.02	0.20	938	1.00	0.13
	7	904	1.01	0.21	904	1.00	0.13
	8	1,001	1.04	0.31	1,001	1.02	0.16
Reading	K	1,066	0.97	0.15	1,066	0.99	0.16
	1	1,037	0.98	0.27	1,037	0.98	0.15
	2	1,114	1.04	0.89	1,114	0.99	0.17
	3	1,102	1.02	0.18	1,102	1.03	0.19
	4	1,183	1.02	0.18	1,183	1.03	0.21
	5	1,219	1.01	0.17	1,219	1.02	0.18
	6	912	1.01	0.17	912	1.02	0.19
	7	892	1.00	0.15	892	1.01	0.21
	8	971	1.01	0.18	971	1.02	0.23

2.5.2 Item fit

The differences between observed and expected item scores were similar on average for both groups. The most noticeable difference in As seen with person outfit, the fit statistics in Table 2.9 were all below one on average. Items are also showing overfitting. Item outfit statistics for the two groups are quite similar.

Table 2.8 was that the pilot group showed more variabilities in the difference. The pilot group's standard deviations were larger than the comparison group's, especially for math. The direction of the difference was also different for the two groups. For math, the pilot group had larger expected scores across all grades whereas the comparison group had a mix. Some comparison group grades had lower expected value and other had larger. In reading, the direction was mixed for both groups.

As seen with person outfit, the fit statistics in Table 2.9 were all below one on average. Items are also showing overfitting. Item outfit statistics for the two groups are quite similar.

Table 2.8. Summary of the difference in observed and expected proportion correct scores for items

		Pilot			Comparison		
Term	Grade	N Items	Mean	S.D.	N	Mean	S.D.
Math	K	207	0.02	0.10	314	0.02	0.10
	1	272	-0.01	0.10	555	0.01	0.09
	2	299	-0.01	0.11	795	-0.02	0.10
	3	559	0.02	0.15	1,208	0.01	0.11
	4	515	0.00	0.13	743	0.00	0.11
	5	378	0.02	0.15	502	0.00	0.11
	6	434	-0.02	0.15	539	-0.03	0.13
	7	316	0.00	0.11	342	0.00	0.11
	8	288	0.02	0.14	224	0.00	0.10
Reading	K	366	0.01	0.10	399	-0.01	0.09
	1	454	0.00	0.09	762	0.00	0.08
	2	473	0.00	0.08	666	0.00	0.08
	3	201	0.01	0.09	564	0.01	0.08
	4	245	0.00	0.09	357	0.01	0.09
	5	211	0.00	0.08	207	0.02	0.09
	6	246	0.00	0.08	232	0.01	0.08
	7	191	0.00	0.09	165	0.00	0.08
	8	171	0.00	0.08	141	0.00	0.08

Table 2.9. Summary of item outfit statistics

		Pilot			Comparison		
Term	Grade	N Items	Mean	S.D.	N	Mean	S.D.
Math	K	207	1.00	0.23	314	1.01	0.26
	1	272	1.16	2.31	555	0.97	0.13
	2	299	1.06	0.41	795	0.99	0.13
	3	559	0.98	0.16	1,208	0.97	0.12
	4	515	1.01	0.18	743	1.00	0.12
	5	378	1.01	0.22	502	1.00	0.13
	6	434	1.01	0.16	539	1.02	0.14
	7	316	1.01	0.18	342	1.01	0.12
	8	288	1.04	0.22	224	1.02	0.13
Reading	K	366	0.98	0.12	399	1.00	0.17
	1	454	1.00	0.34	762	0.99	0.13
	2	473	0.99	0.17	666	1.00	0.13
	3	201	1.02	0.11	564	1.01	0.14
	4	245	1.01	0.11	357	1.02	0.14
	5	211	1.01	0.16	207	1.01	0.11
	6	246	0.99	0.11	232	1.01	0.11
	7	191	0.99	0.12	165	1.02	0.12
	8	171	0.99	0.11	141	1.01	0.13

3. Conclusion

The Content Proximity Project was designed to improve the content validity of the MAP® Growth™ assessments while retaining the ability for the test to adapt off-grade and meet students wherever they are in their learning.

Content Proximity Tests were designed to give greater preference to on-grade items and have different numbers of items per instructional area when necessary. Content Proximity math tests also emphasize mathematics Aspects of Rigor, and reading tests have more reading passages at the upper grade levels. These changes are an evolution of MAP Growth Assessments that allow for better alignment to curriculum and instruction while retaining the ability for a test to adapt off-grade and meet students wherever they may be in their learning. The changes are a refinement of MAP Growth. They are not considered to be very drastic or substantial enough to for the tests to represent new products.

A key finding of the Content Proximity Pilot Study is that content validity is enhanced in math and reading. Test events contain substantially more on-grade items. Off-grade adaptation continues to occur for low- and high-performing students. Moreover, the more extreme the student performance, the more off-grade adaptation occurs.

Another key finding is that Content Proximity reading test scores were comparable to traditional MAP Growth reading scores. Content validity improved with little to no change to reading scores. On the other hand, Content Proximity math test scores were consistently higher than traditional MAP Growth math test scores. The increase varied by grade, but Content Proximity math test scores were about 3 RIT score points higher. The norms predictions for math were different by a similar amount.

4. References

Cheng, Y., & Chang, H-. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369 – 383.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kingsbury, G. G., & Zara, A. R. (1989). Procedure for selecting items for computerizes adaptive tests. *Applied Measurement in Education*, *2*(4), 359 – 375.

Nurakhmetov, D. (2019). Reinforcement learning applied to adaptive classification testing. In B. P. Veldkamp, & C. Sluijter (Eds.) *Theoretical and practical advances in computer-based educational measurement* (pp. 325-336). New York, Springer. https://doi.org/10.1007/978-3-030-18480-3

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38.

Thum, Y. M., & Kingsbury, G. (2017, August). *Some new evidence that MAP Growth scores are on a cross-grade vertical, interval scale*. NWEA White Paper. Portland, OR: NWEA.

Thum, Y. M., & Kuhfeld, M. (2020). NWEA 2020 MAP Growth Achievement Status and Growth Norms for Students and Schools. NWEA Research Report. Portland, OR: NWEA.