Predicting Amira Reading Mastery Based on NWEA MAP Reading Fluency Benchmark Assessment Scores

June 2024

NWEA Psychometrics and Analytics

Linking Study Updates

Date	Description
2024-05	Initial linking study conducted for MAP Reading Fluency and Amira Reading Mastery scores for English grades 1–5 using fall and winter data from the 2023–2024 school year

Acknowledgements: This report was made possible with the contributions of Fang Peng, Ann Hu, and Christopher Wells. We appreciate our colleagues at NWEA and the psychometric team at Amira who provided data for the study.

Table of Contents

List of Tables

1. Introduction

1.1. Purpose of the Study

NWEA® is committed to providing partners with useful tools to help make inferences about student learning from MAP® Reading Fluency™ test scores. With the release of MAP® Reading Fluency Coach,™ students can receive personalized reading tutoring solutions based on their benchmark test scores in order to improve their reading growth.

This document presents results from a linking study conducted by NWEA in May 2024 to statistically connect the grades 1–5 English Amira Reading Mastery (ARM) scores with the Scaled-Words-Correct-Per-Minute (SWCPM) scores from the MAP Reading Fluency benchmark assessment taken during Fall and Winter 2023–2024. In this study, the corresponding MAP Reading Fluency score ranges are also provided for each grade and term so that educators can identify students' performance categories based on their ARM score. This report presents the following results:

1. Student sample demographics
2. Descriptive statistics of test scores
3. Overview of the score-linking procedure
4. MAP Reading Fluency SWCPM score ranges corresponding to the performance categories defined by the ARM score
5. Classification accuracy statistics to determine the degree to which the MAP Reading Fluency score accurately predicts student proficiency based on the ARM score

1.2. Overview of Scores

MAP Reading Fluency SWCPM is an equated score obtained from oral reading tests included in the MAP Reading Fluency benchmark assessment. NWEA routinely conducts equating to place the raw Words-Correct-Per-Minute scores from oral reading passages onto the same scale as that of a reference passage. The resulting adjusted SWCPM score accounts for differences in passage difficulty and can be used to meaningfully compare a student's oral reading performance across tests. SWCPM scores are reported on a vertical scale with a range of 0–170.

An ARM score is a grade-equivalent score derived from Amira's benchmark assessment. It provides an overall assessment of a student's reading mastery and serves as the basis for determining tutoring placement in MAP Reading Fluency Coach.™ For each ARM score, its integer value represents grade-level equivalence, and the decimal value represents the number of months of instruction received. For instance, an ARM score of 1.2 indicates that the student's reading mastery is equivalent to a first grader having received two months of reading instruction in the school year. An ARM score is reported on the same scale within each grade, and its 25th and 75th percentiles are used to define three performance categories based on the ARM score for each grade and term: *Below Grade Level*, *On Grade Level*, and *Above Grade Level*.

In MAP Reading Fluency Coach,™ students' ARM scores are predicted using their MAP Reading Fluency benchmark test scores in order to adaptively assign the most appropriate tutoring content based on their reading progression.

2. Methods

2.1. Data Collection

This linking study is based on data from the Fall and Winter 2023–2024 administrations of the MAP Reading Fluency and Amira assessments. Testing records were extracted for students who completed either a MAP Reading Fluency assessment or an Amira assessment during the target terms. As of this linking study, the Spring 2023–2024 test administration is still in progress. Results for the spring term will be appended to this report once data collection is completed.

2.2. Statistical Matching of Test Records

The testing records for each assessment were initially matched based on common student identifiers to create a dataset comprising students who took both the MAP Reading Fluency and the Amira assessments. However, the resulting dataset did not provide sufficient sample sizes for the linking study and did not accurately represent the demographics and score ranges found in the MAP Reading Fluency population data. To address these limitations, statistical matching methods were utilized to expand the study sample. If a student had taken the MAP Reading Fluency test in a target term but lacked a corresponding Amira record, statistical matching was performed to identify an Amira record from a student with similar characteristics in terms of gender, ethnicity, English language learner status, special education status, and score percentile ranking. Additionally, school characteristics—such as demographic composition, private school status, and urbanicity—were also included in the statistical-matching process.

The goal of statistical matching was to obtain robust linking study data that are representative of the score distributions across grades and test terms. These matched student pairs and their observed test records could then be used to augment the linking study data for subsequent analyses.

A procedural statistical-matching process was established. First, two pools of students were formed for each grade and term: one comprising students who took only the MAP Reading Fluency assessment and their first test records and the other with students who took only the Amira assessment and their first test records. All attribute variables were standardized for uniformity. Euclidean distance was calculated for each pair of students from the two pools to create a distance matrix capturing pairwise similarities based on student-level and school-level attributes. Next, to achieve the one-to-one bipartite matching between the MAP Reading Fluency and Amira student pools, the Kuhn-Munkres algorithm (Kuhn, 1955) was applied to identify the optimal pairing assignments that minimize the sum of distances between all pairs. This process ensures that the global distance is minimized while maximizing local pairwise similarities. Due to smaller sample sizes in the Amira pool for most grades and terms, some MAP Reading Fluency test records were dropped because they could not obtain a match before exhausting the available Amira candidates.

The data were further inspected through scatterplots of MAP Reading Fluency SWCPM scores and ARM scores across different test terms and grades. Noticeable outliers were found at both extremes of the scales. To mitigate the impact of these outliers, the Mahalanobis distance was used to exclude observations that significantly deviate from the bivariate score distribution with the following formula:

$$D^2 = (X - \mu)^T COV^{-1}(X - \mu)$$

where D^2 is the squared Mahalanobis distance of a data point X to the centroid μ of the data distribution, with COV representing the covariance matrix. The squared Mahalanobis distance follows a chi-square distribution with degrees of freedom equal to the number of variables. A threshold value of 13.8 was used, as data points exceeding this critical value are not considered because their probability of occurrence is less than 0.1 under the assumed distribution.

The final linking study sample consisted of students who completed both assessments and had their test scores merged (“common student sample”), as well as students who had only completed the MAP Reading Fluency assessment but were successfully matched with a counterpart who had an available Amira test score (“statistically matched sample”).

2.3. Post-Stratification Weighting

Post-stratification was performed to ensure that the linking study sample represented the gender and ethnicity demographics within the MAP Reading Fluency population. These demographic variables were selected because they are correlated with the student’s test scores and are typically subgroups of interest for generalizing the study findings. Specifically, an iterative raking procedure was used to calculate the post-stratification weights that align sample marginal distributions to known population margins. The weighted sample matches the target MAP Reading Fluency population as closely as possible in key demographics. The following steps were taken during this process:

- Calculate marginal distributions of gender and ethnicity for the sample and population.
- Compute post-stratification weights with the raking function (Lumley, 2019).
- Apply the weights to the sample when conducting the linking study analyses.

2.4. Descriptive Statistics

Descriptive statistics are provided to summarize the test scores for both the MAP Reading Fluency and Amira assessments, including the test score mean, standard deviation (SD), minimum, and maximum. The mean presents the average test scores across all students in the study sample, and the SD indicates the variability of test scores, revealing how students’ scores are distributed around the average score. Correlation coefficients between the two test scores are also calculated to examine the strength of the association. The correlations were calculated as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

2.5. Score Prediction

A linear regression model was used to predict the ARM score based on the MAP Reading Fluency SWCPM score. Given a MAP Reading Fluency SWCPM score of x, its predicted ARM score, y, can be calculated as:

$$y = \beta_0 + \beta_1 x$$

where β_0 is the intercept term, and β_1 is the regression coefficient applied to score x.

Three statistical methods were initially compared for the linking task: equipercentile equating, mean-sigma linear equating (Kolen & Brennan, 2004), and linear regression. The performance of these models was evaluated using the Root-Mean-Square Error (RMSE) of the predictions, calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents an observed ARM score, $\hat{y}_i$ is the corresponding predicted score, and n is the number of observations. Linear regression outperformed the other models and was therefore selected as the final model.

Data from both the common student sample and the statistically matched sample were combined to construct the prediction model with weighted least square estimation. However, only the common student data were used for validation and model selection. The model's performance was evaluated exclusively on the common student data, ensuring that the final model was chosen based on its prediction accuracy for genuinely matching test records.

2.6. Classification Accuracy

To assist educators in interpreting the ARM score, three performance levels were defined based on normative data: *Below Grade Level*, *On Grade Level*, and *Above Grade Level*. These levels were established using the 25^{th} and 75^{th} percentiles of the ARM scores. The 25^{th} percentile of the ARM scores marks the boundary between *Below Grade Level* and *On Grade Level*, while the 75^{th} percentile distinguishes *On Grade Level* from *Above Grade Level*.

To measure classification agreement, the percentages of exact match and Cohen's kappa values were calculated. The exact match rate shows the proportion of instances where the predicted class exactly matches the true class. While this measure is easy to understand, it does not account for the possibility of agreement occurring by chance. Cohen's kappa, on the other hand, is a more robust measure because it describes the level of agreement beyond chance. It is calculated as:

$$Kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the exact match rate, and p_e is the sum of the products of marginal totals (i.e., the i^{th} row total multiplied by the i^{th} column total of the classification contingency table) divided by the square of the total number of ratings. Kappa values tend to be lower than percent agreement but are interpreted differently. For instance, a kappa value of 0.2 indicates a 20% improvement over what would be expected by chance.

The linking study results are further evaluated by checking how accurately the predicted ARM scores classify students into two categories: proficient (*On Grade Level* and *Above Level*) or not proficient (*Below Grade Level*). Table 2.1 describes the classification accuracy statistics considered in this report (Pommerich et al., 2004).

Table 2.1. Description of Classification Accuracy Summary Statistics

Statistic	Description	Interpretation
Overall Classification Accuracy Rate	(TP + TN) / (total sample size)	Proportion of the study sample whose proficiency classification based on the ARM score was correctly predicted by MAP Reading Fluency SWCPM
False Negative (FN) Rate	FN / (FN + TP)	Proportion of not-proficient students identified by MAP Reading Fluency SWCPM in those observed as proficient based on the ARM score
False Positive (FP) Rate	FP / (FP + TN)	Proportion of proficient students identified by MAP Reading Fluency SWCPM in those observed as not-proficient based on the ARM score
Sensitivity	TP / (TP + FN)	Proportion of proficient students identified by MAP Reading Fluency SWCPM in those observed as such based on the ARM score
Specificity	TN / (TN + FP)	Proportion of not-proficient students identified by MAP Reading Fluency SWCPM in those observed as such based on the ARM score
Precision	TP / (TP + FP)	Proportion of observed proficient students based on the ARM score in those identified as such by the MAP Reading Fluency SWCPM
Area Under the Curve (AUC)	Area under the receiver operating characteristics (ROC) curve	How well MAP Reading Fluency SWCPM scores distinguish the study sample into proficiency categories that match those from the ARM scores. To determine this, a logistic regression was performed, predicting the ARM score proficiency category using the SWCPM score. Students' predicted proficiency probabilities were compared with their actual ARM proficiency categories to calculate the AUC score. An AUC at or above 0.80 is considered "good" accuracy.

Note. FP = false positives; FN = false negatives; TP = true positives; TN = true negatives.

3. Results

3.1. Study Sample

The linking study sample includes data from the common student sample and the statistically matched sample. In the statistically matched sample, MAP Reading Fluency students were paired with their counterparts in the Amira student sample based on the distance measure of matching student attributes. Table 3.1 presents the data distributions of matching variables in both student pools. Due to the differences in population distribution and the multi-variate distant-matching method, MAP Reading Fluency students could not always obtain a counterpart with exact match across all attributes. However, the marginal distributions across the demographics are reasonably close between the MAP Reading Fluency students and the Amira students in the final linking study sample.

Table 3.1. Distribution of Matching Variables in Linking Study Sample by Student Pool

Matching Attributes			Grade 1	Grade 2	Grade 3	Grade 4	Grade 5
MAP Reading Fluency Student Pool							
		Total N	18,256	23,718	23,345	8,797	7,534
Student-Level Attributes	Score Percentile	Avg. SWCPM Percentile	49.3	53.2	49.4	46.5	46.3
	Ethnicity	White (%)	43.2	40.8	38.8	35.8	31.9
		Hispanic (%)	21.8	19.1	20.5	29.1	30.8
		Black (%)	16.2	26.3	26.5	20.3	22.7
		Asian (%)	7.7	4.3	4.8	3.3	3.6
		Multi-Race (%)	11.1	9.6	9.4	11.5	11.0
	Gender	Female (%)	50.0	50.3	49.7	51.3	49.6
		Male (%)	50.0	49.7	50.3	48.7	50.4
	Other	Special Education (%)	1.3	1.4	1.6	1.6	1.0
		English Language Learner (%)	2.4	1.2	1.4	1.8	1.8
School-Level Attributes	Locale	City (%)	31.8	20.6	21.8	33.0	37.4
		Suburb (%)	46.3	58.7	58.5	41.1	39.1
		Town (%)	7.6	6.3	6.1	14.2	12.9
		Rural (%)	14.3	14.4	13.7	11.8	10.7
	School Type	Public School (%)	93.2	98.7	98.7	100.0	100.0
		Private School (%)	6.8	1.3	1.3	0.0	0.0
	Ethnicity Composition	Avg. % White	47.9	46.6	44.9	39.9	37.6
		Avg. % Hispanic	14.8	20.1	21.2	20.3	21.3
		Avg. % Black	25.9	21.8	22.5	28.4	29.1
		Avg. % Asian	5.5	4.3	4.0	2.8	3.0
		Avg. % Multi-Race	5.9	7.1	7.4	8.6	8.8
	Gender Composition	Avg. % Female	49.8	49.1	48.8	48.8	48.7
		Avg. % Male	51.1	51.1	51.4	51.2	51.3
Amira Student Pool							
Student-Level Attributes	Score Percentile	Avg. ARM Percentile	49.5	55.3	49.7	46.0	44.7
	Ethnicity	White (%)	39.7	40.8	38.8	35.8	31.9
		Hispanic (%)	22.0	19.1	20.5	29.1	30.8
		Black (%)	21.7	26.3	26.6	20.3	22.7

Matching Attributes			Grade 1	2	3	4	5
		Asian (%)	6.1	4.3	4.8	3.3	3.6
		Multi-Race (%)	10.4	9.6	9.3	11.5	11.0
	Gender	Female (%)	49.9	50.3	49.7	51.2	49.6
		Male (%)	50.1	49.7	50.3	48.8	50.4
	Other	Special Education (%)	1.5	1.4	1.4	1.8	1.0
		English Language Learner (%)	1.9	1.1	1.1	1.8	1.8
School-Level Attributes	Locale	City (%)	24.6	18.1	17.4	32.9	37.4
		Suburb (%)	53.0	61.2	62.9	41.1	39.1
		Town (%)	8.2	6.3	6.1	14.2	12.9
		Rural (%)	14.2	14.4	13.6	11.8	10.7
	School Type	Public School (%)	99.6	99.5	99.5	100.0	100.0
		Private School (%)	0.4	0.5	0.5	0.0	0.0
	Ethnicity Composition	Avg. % White	42.2	42.3	41.0	38.8	36.3
		Avg. % Hispanic	25.1	26.5	26.4	21.2	22.4
		Avg. % Black	20.0	19.3	20.5	28.1	28.8
		Avg. % Asian	4.7	4.6	4.7	3.3	3.4
		Avg. % Multi-Race	7.9	7.2%	7.2	8.6	9.0
	Gender Composition	Avg. % Female	48.5	48.5	48.6	48.6	48.5
		Avg. % Male	51.6	51.6	51.5	51.4	51.5

The final MAP Reading Fluency sample includes students from 49 states, 884 districts, and 2,803 schools in the United States. Table 2 presents the numbers of students by grade and the percentages by U.S. region within each test window. The study sample includes students from all four U.S. regions.

Table 2.2. Numbers and Percentages of Students by U.S. Region

Grade	N	Percentage (%) Northeast	Midwest	South	West
Fall					
1	7,846	7.0	31.7	54.3	7.0
2	12,622	5.7	37.4	51.5	5.4
3	12,256	6.1	40.8	47.2	6.0
4	4,916	2.8	22.5	65.8	8.8
5	4,432	3.5	19.9	67.7	9
Winter					
1	10,410	9.0	36.1	46.7	8.2
2	11,096	5.5	36.7	52.3	5.4
3	11,089	7.1	43.1	45.2	4.6
4	3,881	4.4	26.2	60.1	9.3
5	3,102	4.1	29.8	56.0	10.1

Since the unweighted MAP Reading Fluency sample distribution is different from the student population, post-stratification weights were applied to the linking study sample to improve its representativeness. Table 3.3 presents the demographic distributions of gender and ethnicity in the target MAP Reading Fluency student population by grade and term. Table 3.4 and Table 3.5 present demographics of the MAP Reading Fluency student sample in the final linking study data before and after post-stratification weights were applied, respectively. The weighted sample distributions are almost identical to the MAP Reading Fluency student population distributions. The analyses in this study were therefore conducted based on the weighted sample.

Table 3.3. MAP Reading Fluency Student Population Demographics, 2023–2024

Demographic Subgroup		% Students by Grade				
		1	2	3	4	5
Fall						
	Total N	8,105	54,934	62,711	43,732	34,356
Ethnicity	White	42.2	50.3	49.3	43.5	42.7
	Hispanic	23.4	21.5	22.9	28.6	28.7
	Black	11.7	11.4	11.7	13.5	14.4
	Asian	10.0	5.7	4.4	2.9	3.1
	Multi-Race	12.7	11.2	11.7	11.5	11.2
Gender	Female	49.8	52.0	51.3	50.6	50.3
	Male	50.2	48.0	48.7	49.4	49.7
Winter						
	Total N	18,633	67,291	60,171	37,438	28,585
Ethnicity	White	49.0	50.2	47.4	42.6	41.9
	Hispanic	19.7	22.1	24.4	30.3	30.3
	Black	11.9	11.7	12.8	13.3	14.5
	Asian	7.7	5.2	4.2	2.9	3.1
	Multi-Race	11.6	10.9	11.1	11.0	10.2
Gender	Female	51.1	52.1	51.4	50.6	49.9
	Male	48.9	47.9	48.6	49.4	50.1

Table 3.4. MAP Reading Fluency Linking Study Student Sample Demographics (Unweighted), 2023–2024

Demographic Subgroup		% Students by Grade				
		1	2	3	4	5
Fall						
	Total N	7,846	12,622	12,256	4,916	4,432
Ethnicity	White	42.3	41.9	38.4	34.8	30.9
	Hispanic	23.3	18.1	20.7	28.8	30.5
	Black	11.8	25.6	26.2	21.0	23.1
	Asian	10.0	4.0	4.7	2.9	3.4
	Multi-Race	12.7	10.4	10.0	12.6	12.1
Gender	Female	49.9	50.5	49.7	51.3	49.8
	Male	50.1	49.5	50.3	48.7	50.2
Winter						
	Total N	10,410	11,096	11,089	3,881	3,102
Ethnicity	White	44.0	39.5	39.2	37.0	33.2

Demographic Subgroup		% Students by Grade				
		1	2	3	4	5
	Hispanic	20.7	20.1	20.3	29.5	31.3
	Black	19.5	27.0	20.3	19.5	22.2
	Asian	6.0	4.7	4.9	3.8	3.8
	Multi-Race	9.9	8.7	8.7	10.2	9.4
Gender	Female	50.0	50.1	49.6	51.3	49.5
	Male	50.0	49.9	50.4	48.7	50.5

Table 3.5. MAP Reading Fluency Linking Study Student Sample Demographics (Weighted), 2023–2024

Demographic Subgroup		% Students by Grade				
		1	2	3	4	5
Fall						
	Total N	7,846	12,622	12,256	4,916	4,432
Ethnicity	White	42.2	50.3	49.3	43.5	42.7
	Hispanic	23.4	21.5	22.9	28.6	28.7
	Black	11.7	11.4	11.7	13.5	14.4
	Asian	10.0	5.7	4.4	2.9	3.1
	Multi-Race	12.7	11.2	11.7	11.5	11.2
Gender	Female	49.8	52.0	51.3	50.6	50.3
	Male	50.2	48.0	48.7	49.4	49.7
Winter						
	Total N	10,410	11,096	11,089	3,881	3,102
Ethnicity	White	49.0	50.2	47.4	42.6	41.9
	Hispanic	19.7	22.1	24.4	30.3	30.3
	Black	11.9	11.7	12.8	13.3	14.5
	Asian	7.7	5.2	4.2	2.9	3.1
	Multi-Race	11.6	10.9	11.1	11.0	10.2
Gender	Female	51.1	52.1	51.4	50.6	49.9
	Male	48.9	47.9	48.6	49.4	50.1

3.2. Descriptive Statistics

Table 3.6 presents descriptive statistics of the MAP Reading Fluency SWCPM and the Amira ARM test scores by term and grade, including the correlation coefficient (*r*) between them. The coefficients between the scores range from 0.90 to 0.96. These values indicate a high positive correlation among the scores, which is important validity evidence for the claim that the MAP Reading Fluency SWCPM score is a good predictor of the ARM score.

Table 3.6. Descriptive Statistics of Test Scores

Grade	Term	N	*r*	MAP Reading Fluency SWCPM				ARM			
				Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
1	Fall	7,846	0.90	75.57	22.80	11	150	1.06	0.63	0.12	2.57
	Winter	10,410	0.91	77.51	22.29	10	149	1.45	0.63	0.28	2.79
2	Fall	12,622	0.96	83.20	23.51	10	154	2.22	0.81	0.13	4.33

Grade	Term	N	*r*	MAP Reading Fluency SWCPM				ARM			
				Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
	Winter	11,096	0.95	92.89	24.01	26	170	2.61	0.75	0.60	4.78
3	Fall	12,256	0.96	95.08	26.14	19	170	2.90	1.01	0.45	5.78
	Winter	11,089	0.96	103.65	26.56	23	170	3.46	0.91	0.80	5.78
4	Fall	4,916	0.93	100.16	28.65	4	170	3.64	1.06	0.59	6.60
	Winter	3,881	0.95	107.66	26.22	25	170	4.10	1.01	0.76	6.77
5	Fall	4,432	0.91	118.38	30.14	13	170	4.74	1.06	1.02	7.06
	Winter	3,102	0.94	121.09	28.52	32	170	5.11	0.92	1.88	7.39

Note. SD = standard deviation; Min. = minimum; Max. = maximum.

3.3. Score Prediction Model

Table 3.7 presents the Root-Mean-Square-Error (RMSE) values across three linking methods calculated based on the common student sample. Overall, the differences in performance are minimal across the models. The models' performance is expected to be lower for the first-grade data due to the narrower range of ARM scores compared with other grades, resulting in reduced score variability. This limitation restricts the models' ability to detect differences in ARM scores or associations between test scores. Although the equipercentile method showed slightly better performance for grade 1 data, the linear regression model consistently outperformed all other models for the remaining grades and was therefore selected as the final prediction model.

Table 3.7. Model Performance Comparison Based on Root-Mean-Square-Error

Linking Method	Grade				
	1	2	3	4	5
Fall					
Equipercentile	0.61	0.51	0.40	0.38	0.43
Mean-Sigma Linear Equating	0.64	0.51	0.40	0.37	0.42
Linear Regression	0.64	0.50	0.39	0.35	0.39
Winter					
Equipercentile	0.68	0.50	0.39	0.35	0.40
Mean-Sigma Linear Equating	0.75	0.49	0.37	0.34	0.37
Linear Regression	0.74	0.47	0.35	0.33	0.36

3.4. Classification Accuracy

Table 3.8 presents ARM score ranges for *Below Grade Level*, *On Grade Level*, and *Above Grade Level* and the corresponding MAP Reading Fluency SWCPM score ranges for each grade and term. Bold values indicate SWCPM scores that are considered to be at least *On Grade Level* per the ARM score. These values can be used to predict a student's likely ARM score performance level when the MAP Reading Fluency oral reading benchmark test is taken in the fall and winter. For example, a third-grade student who achieves an SWCPM score of 84 in the fall is likely to achieve an ARM score of 2.48, predicting this student to be *On Grade Level* (proficient) for that term.

Overall, the score ranges generally show an expected pattern of monotonic progression across the terms within grades, except for the first grade. This anomaly can be attributed to the transitional nature of first-grade reading, where early readers typically start with oral reading of connected text. While some students initiate this transition earlier in the fall, most are expected to read connected text by the end of the first grade, which is reflected by the differences in sample sizes and score variability across terms. As a result, the student samples and score distributions vary significantly between fall and winter for first graders. Additionally, the differences in scaling between the SWCPM and ARM scores, along with prediction errors, could also contribute to the observed non-monotonic progression in first-grade scores.

Table 3.8. MAP Reading Fluency SWCPM for Predicted ARM Score Ranges for Performance Levels

ARM Score Performance Levels			
Grade	***Below Grade Level***	***On Grade Level***	***Above Grade Level***
Fall			
1	0–0.59	0.60–1.65	1.66–2.41
2	0–1.48	1.49–2.66	2.67–4.07
3	0–2.47	2.48–3.79	3.80–5.05
4	0–3.40	3.41–4.56	4.57–5.84
5	0–4.62	4.63–5.74	5.75–7.06
Winter			
1	0–0.78	0.79–1.96	1.97–2.60
2	0–1.78	1.79–3.07	3.08–4.33
3	0–2.83	2.84–4.12	4.13–5.31
4	0–3.82	3.83–5.01	5.02–6.37
5	0–4.88	4.89–5.93	5.94–7.24
Corresponding MAP Reading Fluency Score Ranges			
Grade	***Below Grade Level***	***On Grade Level***	***Above Grade Level***
Fall			
1	0-56	**57**–99	100–170
2	0–60	**61**–97	98–170
3	0–83	**84**–119	120–170
4	0–92	**93**–127	128–170
5	0–114	**115**–150	151–170
Winter			
1	0-51	**52**–98	99–170
2	0–65	**66**–109	110–170
3	0–84	**85**–124	125–170
4	0–99	**100**–133	134–170
5	0–116	**117**–152	153–170

Table 3.9 presents the classification agreement rates and kappa values for the three performance levels based on ARM score. Table 3.10 presents the classification accuracy summary statistics for the two performance categories: proficient and not proficient. These results indicate how well the MAP Reading Fluency SWCPM scores predict the performance levels based on ARM scores (as well as proficiency based on the ARM score), providing insight into the predictive validity of MAP Reading Fluency. The exact match rates range from 74.6 to 90.8. Cohen's kappa values range from 0.60 to 0.85. Both the exact match rates and Cohen's kappa values suggest a moderate to high level of agreement. The overall classification accuracy

rate ranges from 0.81 to 0.96. These values suggest that the SWCPM scores are effective at classifying students into the two ARM score proficiency categories.

Table 3.9. Classification Agreement Rates for Categories of ARM Score

Grade	N	Exact Match (%)	Kappa
Fall			
1	7,846	80.6	0.67
2	12,622	90.8	0.85
3	12,256	84.6	0.75
4	4,916	77.0	0.64
5	4,432	74.6	0.60
Winter			
1	10,410	81.6	0.66
2	11,096	89.6	0.80
3	11,089	89.1	0.82
4	3,881	77.9	0.65
5	3,102	79.6	0.68

Table 3.10. Classification Accuracy Results

Grade	N	Class. Accuracy	Rate		Sensitivity	Specificity	Precision	AUC
			FP	FN				
Fall								
1	7,846	0.89	0.02	0.32	0.68	0.98	0.93	0.96
2	12,622	0.96	0.03	0.09	0.91	0.97	0.86	0.99
3	12,256	0.88	0.15	0.01	0.99	0.85	0.68	0.99
4	4,916	0.82	0.24	0.04	0.96	0.76	0.61	0.96
5	4,432	0.81	0.25	0.07	0.93	0.75	0.63	0.95
Winter								
1	10,410	0.89	0.01	0.50	0.50	0.99	0.95	0.98
2	11,096	0.94	0.04	0.25	0.75	0.96	0.73	0.98
3	11,089	0.94	0.07	0.03	0.97	0.93	0.77	0.99
4	3,881	0.82	0.23	0.01	0.99	0.77	0.57	0.98
5	3,102	0.84	0.20	0.04	0.96	0.80	0.64	0.97

Note. Class. Accuracy = overall classification accuracy rate; FP = false positives; FN = false negatives; AUC = area under the ROC curve.

3.5. Guidelines for Using the Study Results

The sole purpose of this study is to inform MAP Reading Fluency Coach placement by linking the MAP Reading Fluency benchmark score with the ARM score. When students take the Oral Reading Fluency benchmark tests and obtain SWCPM test scores, their likely ARM scores can be predicted accordingly and serve as the basis to select personalized tutoring content tailored to the students' current reading level.

While the correlations between the SWCPM and ARM scores are reasonably strong ($r \geq 0.9$) for all grades and terms, numerous differences exist between the test scores. Some key differences include: 1) the MAP Reading Fluency SWCPM scores are vertically scaled, whereas the ARM scale score metric for each grade is determined independently of those for other grades; 2) the performance levels of SWCPM scores and the performance/proficiency categories of ARM scores are determined using different criteria; 3) the criteria used to make instruction intervention decisions differ. Therefore, it is not recommended to compare the instructional decisions suggested by MAP Reading Fluency scores with those based on the ARM scores.

Although the results show that MAP Reading Fluency SWCPM scores can be effectively used to predict student performance based on the ARM score with relatively high accuracy, there are important limitations to consider regarding the use and interpretation of these results. The method used to link the scores is a one-way predictive model. Therefore, scores on the two tests cannot be assumed to be equivalent or interchangeable. The predicted ARM score is strictly intended to gauge student's approximate reading level so that appropriate tutoring content can be assigned in order to maximize their learning efficacy. A student's predicted ARM score may not be used as a substitute for their MAP Reading Fluency SWCPM score for benchmarking purposes, and their predicted performance category based on their ARM score should not be used to interpret the SWCPM benchmark score or its performance levels. In addition, it is important to recognize that the prediction of ARM scores from MAP Reading Fluency SWCPM scores is not flawless, given the imperfect correlation between them.

References

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking Methods and practices* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics*, *2*(1–2), 83–97. https://doi.org/10.1002/nav.3800020109

Lumley, T. (2019). *Survey: Analysis of complex survey samples* (R package version 3.36) [Computer software]. Available from https://CRAN.R-project.org/package=survey.

Pommerich, M., Hanson, B., Harris, D., & Sconing, J. (2004). Issues in conducting linkage between distinct tests. *Applied Psychological Measurement*, *28*(4), 247–273. https://doi.org/10.1177/0146621604265033