MOBYMAX MATHEMATICS BENCHMARKER ASSESSMENTS AND STATE STANDARDIZED TESTS

A Correlational Validity Study

William Koch

TABLE OF CONTENTS

INTRODUCTION:

In recent years, digital assessments aimed at assessing student growth more effectively and efficiently have become widely available. However, research on the validity of these new assessment products is often lacking, leaving districts, schools, and teachers with inadequate information to make important decisions about which products are best suited to meet their needs. With this in mind, the following research will analyze the validity of one such assessment product: MobyMax's Benchmarker Assessment for mathematics.

Crocker and Algina (1986) define three main types of validity studies: (1) content validity, which establishes the alignment between the test questions and the content it is intended to assess; (2) criterion-related validity, which establishes the relationship between the test score and an outcome or measure that has already been validated and (3) construct validity, which establishes the extent to which the test measures what it purports to measure. This paper will deal with the third type of validity by comparing student scores on Mobymax's Benchmarker Assessments in mathemetics with student scores on a number of state standardized mathematics tests.

Correlations like this are a commonly used and widely accepted form of validity evidence. Correlations demonstrate that when students score high on one assessment, they also tend to score high on the other and vice versa. A high correlation between two assessments provides evidence that the two assessments are measuring similar constructs.

METHODS:

In order to compare MobyMax's Benchmarker Assessments with state standardized tests, MobyMax contacted teachers that were using the program via email and asked them to voluntarily submit anonymized student scores for both Mobymax's Benchmarker Assessments in mathematics and their state standardized mathematics test. These scores were then correlated using the methods described below.

Two statistical methods were used to analyze the data. The main method, Pearson Correlation, is a measure of how closely linked two independent variables are from the same observation (Greene 2002). Correlation describes both the strength and direction of the statistical relationship. Strength indicates how much a change in one variable is also observed in another variable. The stronger the relation between the variable, the higher the correlation coefficient. Direction indicates whether there is a positive relationship (variables move in the same direction) or negative relationship (variable move in the opposite direction). The correlation coefficient is bounded by -1 and 1. Any correlation coefficient with an absolute value > .5 is generally considered a strong statistical relationship depending upon the subject and context (Cohen, 1988).

The second method, Multiple Imputation, is a statistical approach to fill in missing values in a data set based upon the properties of non-missing observations. Missing data is a common problem in statistical analysis. Multiple imputation allows for smaller standard errors and more narrow confidence intervals when performing analyses on imputed data sets. This helps to minimize bias thus improving the accuracy of the predicted value. The particular imputation method used is Multiple Imputation using Chained Equations (MICE) with the Bayesian Linear Regression method more commonly known as the norm method (Chhabra, Vashisht, Ranjan, 2017).

RESULTS:

Table 1 shows all resulting correlations between MobyMax Benchmark Assessments and state standardized tests for which data was received. Any result with a correlation coefficient (r) greater than or equal to 0.5 is considered a strong statistical relation (Cohen, 1988). All the tests in Table 1 have a coefficient greater than 0.5 and are either highly statistically significant at the $p < .01$ level or higher, or in one case, at the $p < 0.1$ level. In this latter result, however, the number of observations was relatively low ($n = 5$) yielding relatively few degrees of freedom. These results indicate that there is a strong correlation between MobyMax Benchmarker Assessments for mathematics and state standardized mathematics tests.

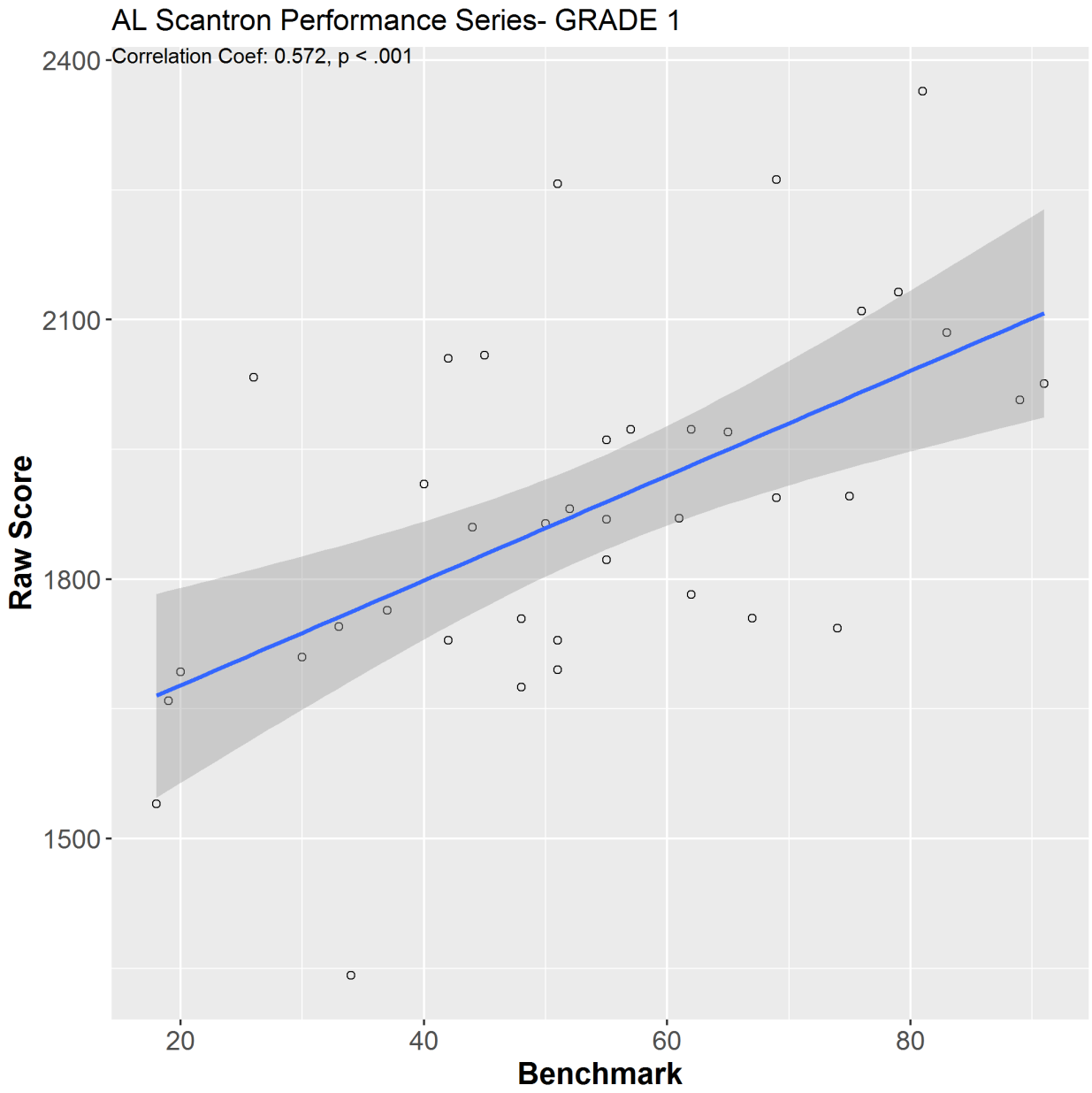
Table 1 - MobyMax: Correlation statistics

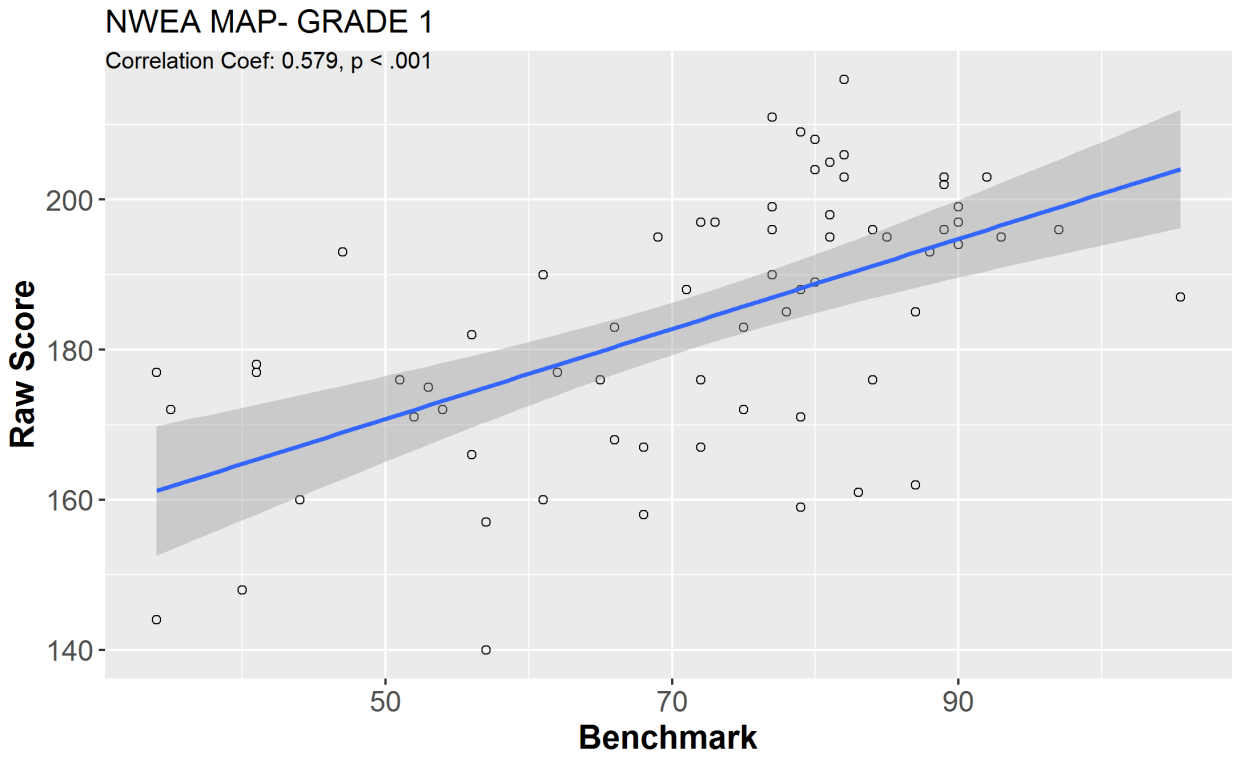
State	*Test*	*Grade Level*	*n**	*r*	r^2	*p-value*
	NWEA MAP	1	67	0.579	0.335	$p < .001$
AL	Scantron Performance Series	1	39	0.572	0.327	$p < .001$
	NWEA MAP	2	21	0.56	0.314	$p = 0.004$
	NWEA MAP	3	55	0.805	0.648	$p < .001$
AL	Scantron Performance Series	3	22	0.611	0.373	$p = 0.001$
SD	Smarter Balanced	3	51	0.784	0.615	$p < .001$
NC	North Carolina End-of-Grade Test	4	21	0.733	0.537	$p < .001$
AL	Scantron Performance Series	5	18	0.762	0.581	$p < .001$
GA	Georgia Milestones Assessment	5	59	0.839	0.704	$p < .001$
PA	Pennsylvania System of School Assessment (PSSA)	5	21	0.809	0.654	$p < .001$
TX	The State of Texas Assessments of Academic Readiness (STAAR)	5	5	0.724	0.524	$p = 0.083$

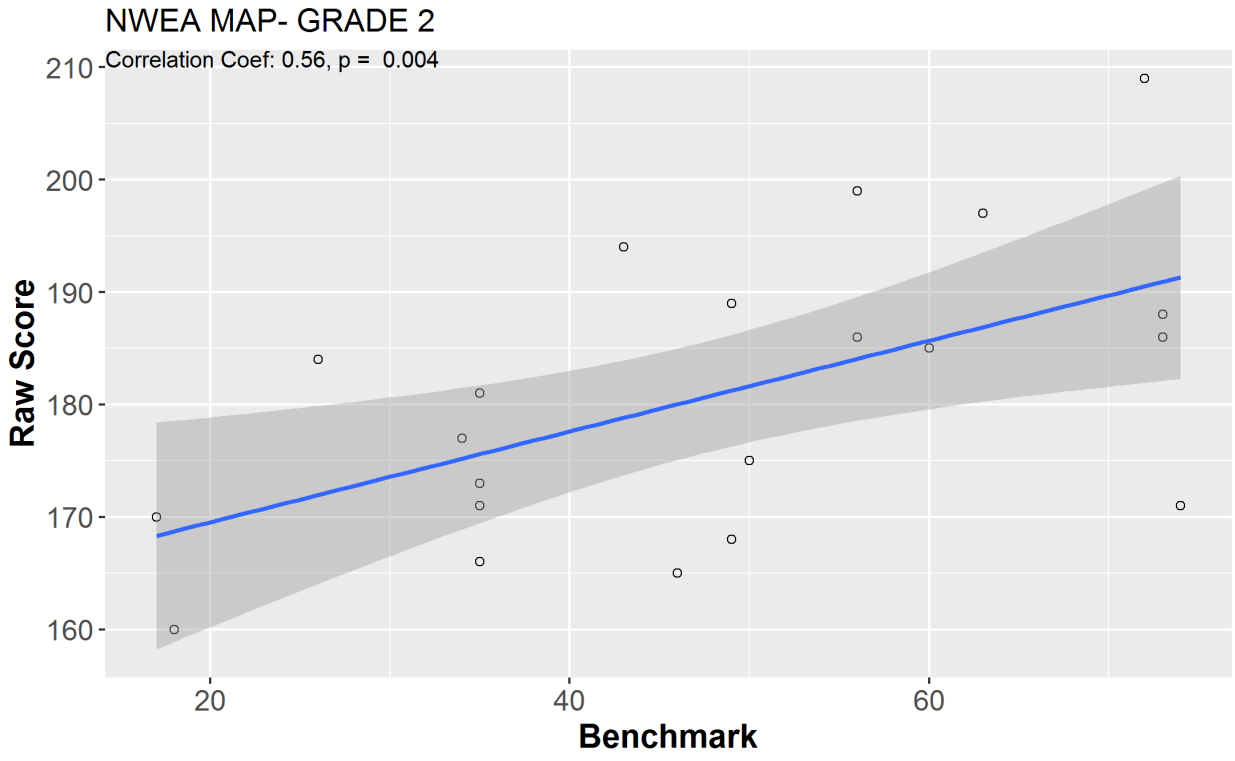
**: Some data were imputed using Multiple Imputation using Chained Equations (MICE) and Bayesian Linear Regression method*

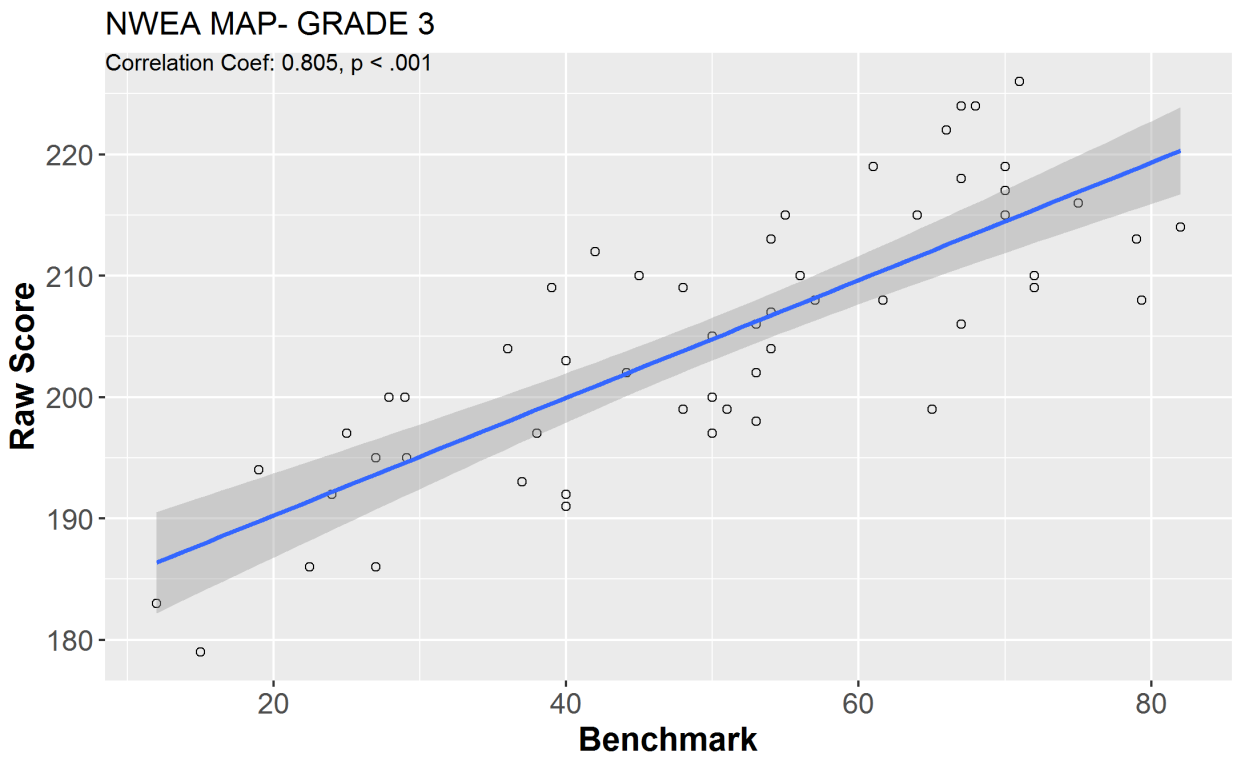
GRAPHS:

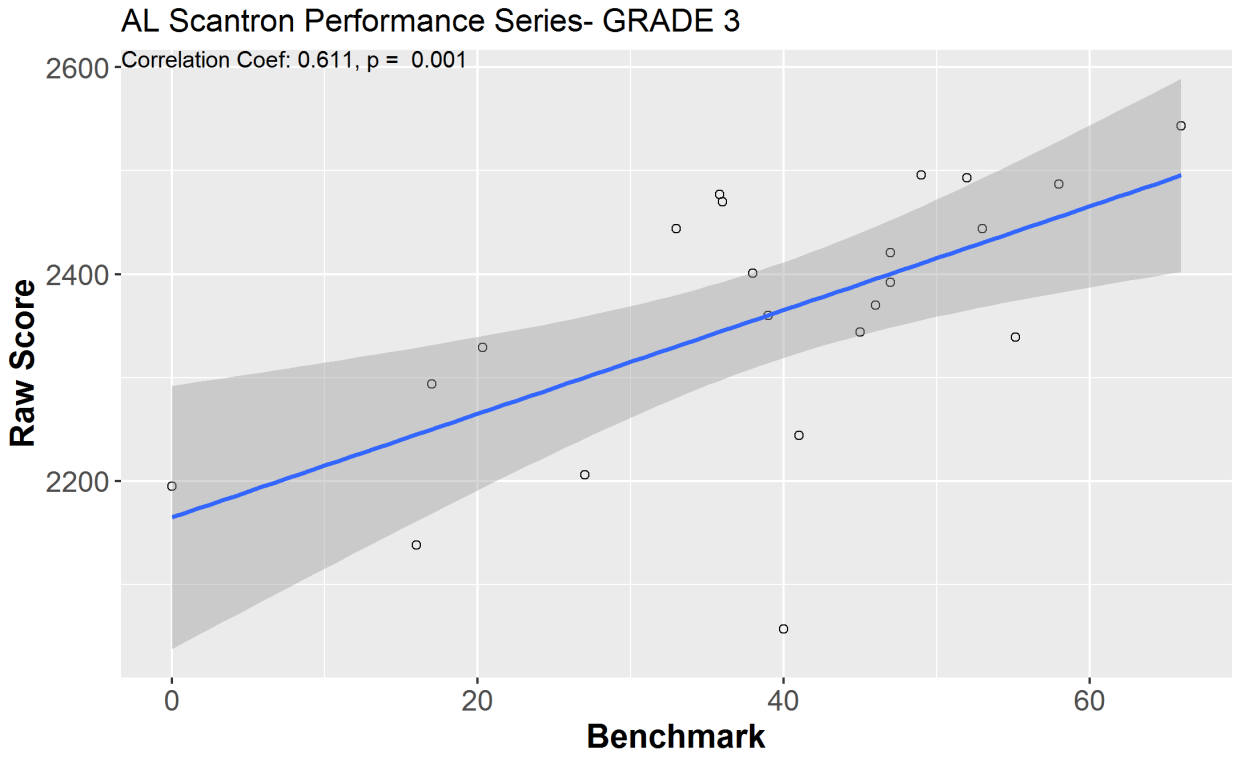
The following graphs are a visual plot of the data for each of the table entries along with the correlation result and statistical significance. The plots are presented according to their order in Table 1. The y-axis "Raw Score" is the unconverted test score for each student on their state standardized math test. The x-axis "Benchmark" is the test score for each student on the MobyMax Benchmarker Assessment for mathematics. The blue line in each graph represents a linear model plot overlay on the original data. The grey area represents a 95% confidence interval around the linear model.

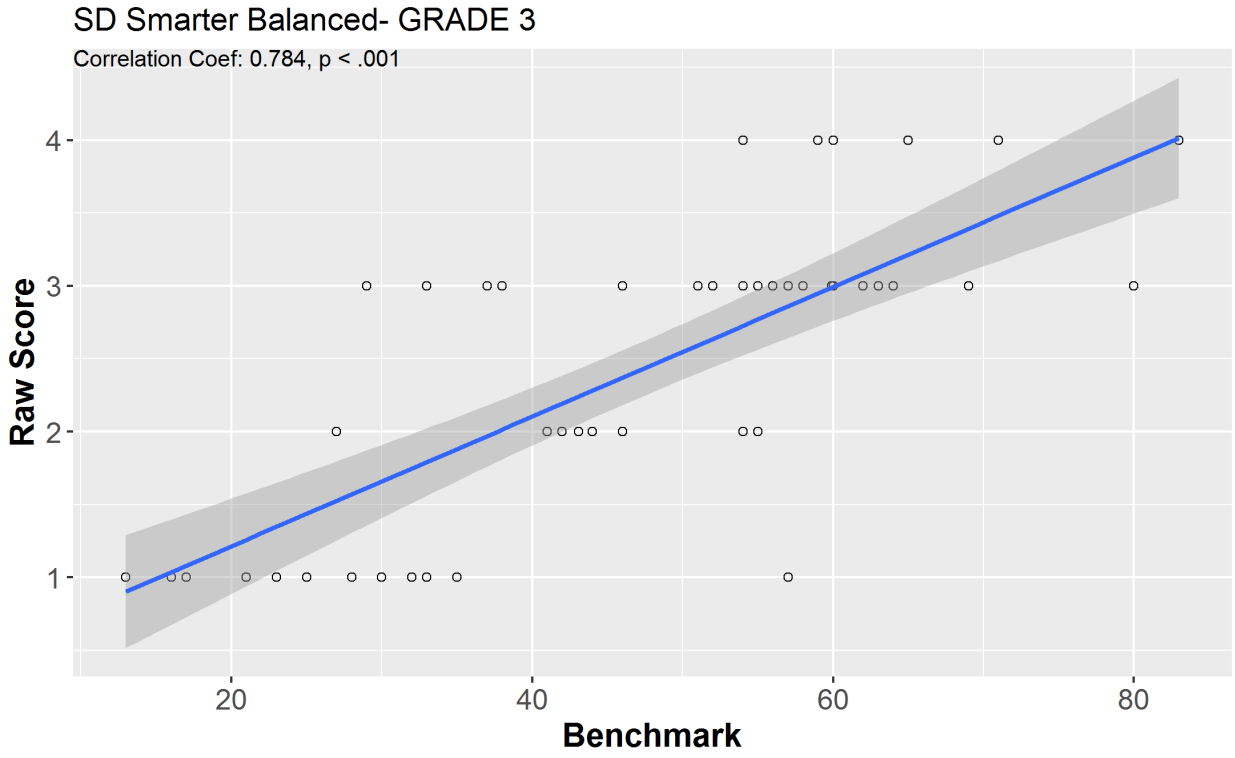


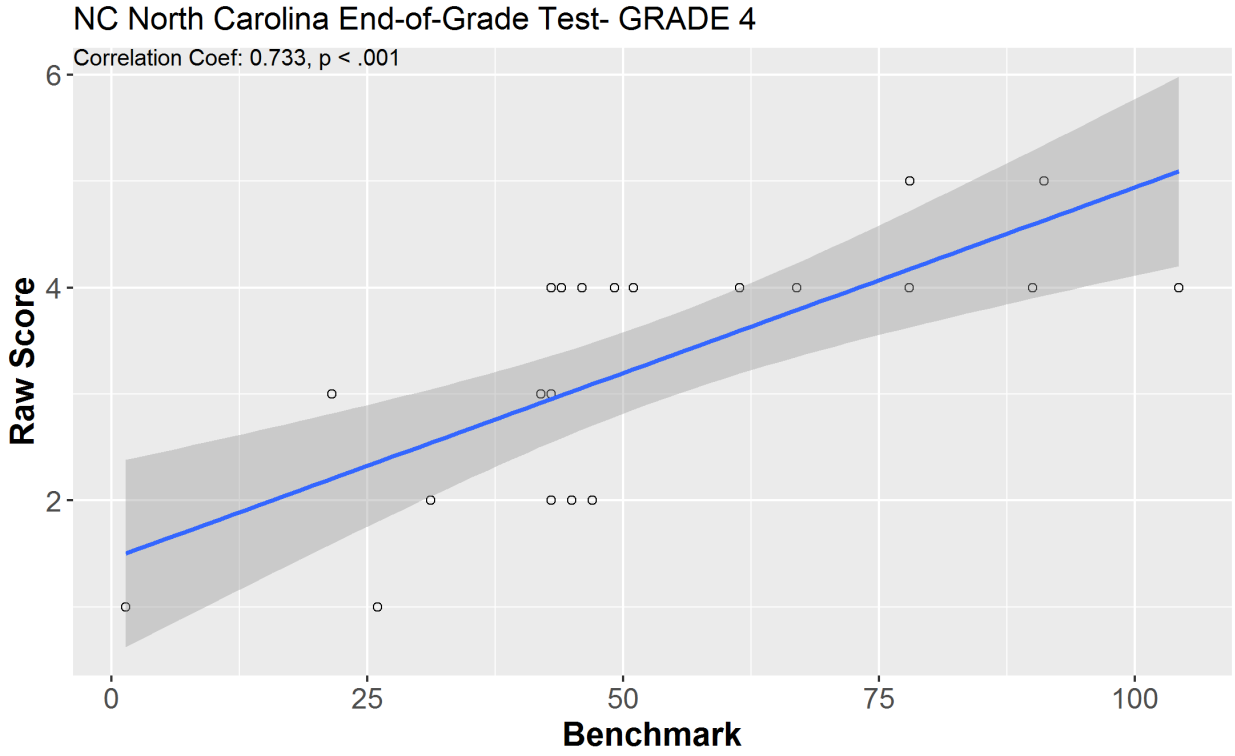

NWEA MAP- GRADE 1
Correlation Coef: 0.579, p < .001
Raw Score
200
180
160
140
50
70
90
Benchmark

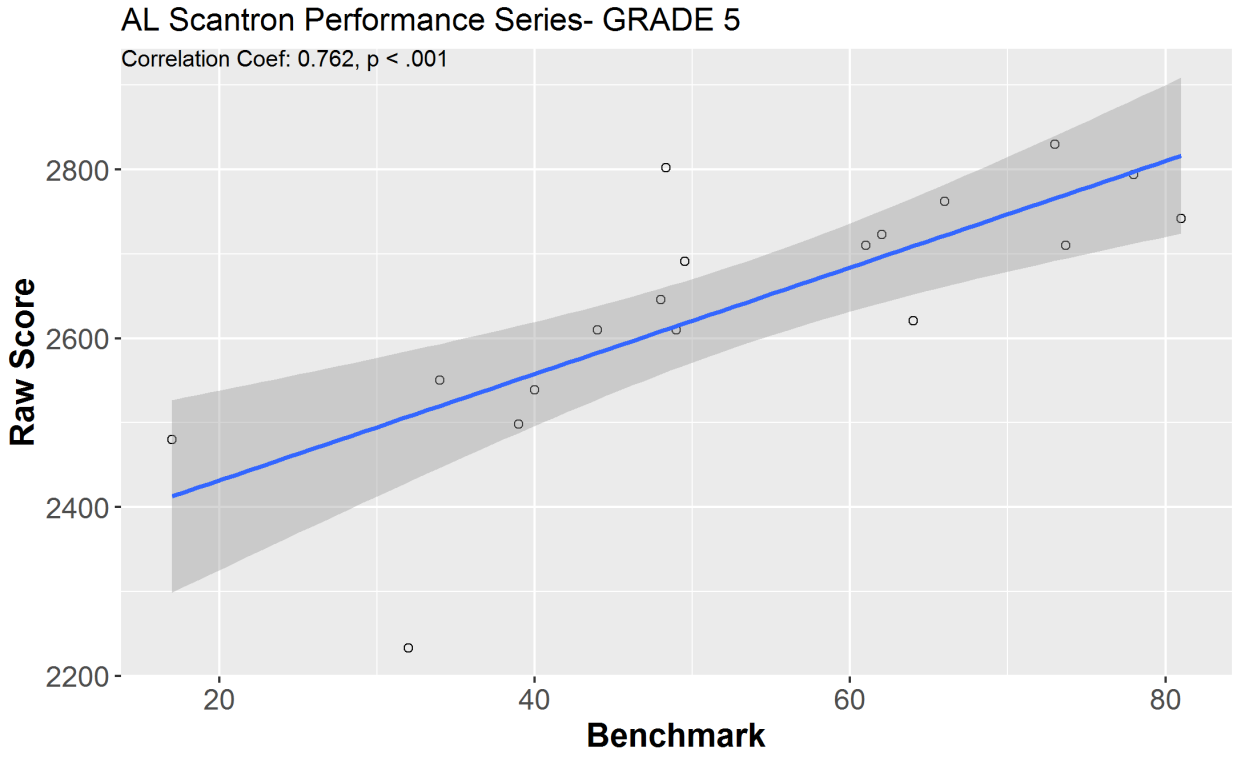

NWEA MAP- GRADE 2
Correlation Coef: 0.56, p = 0.004
Raw Score
210
200
190
180
170
160
20
40
60
Benchmark

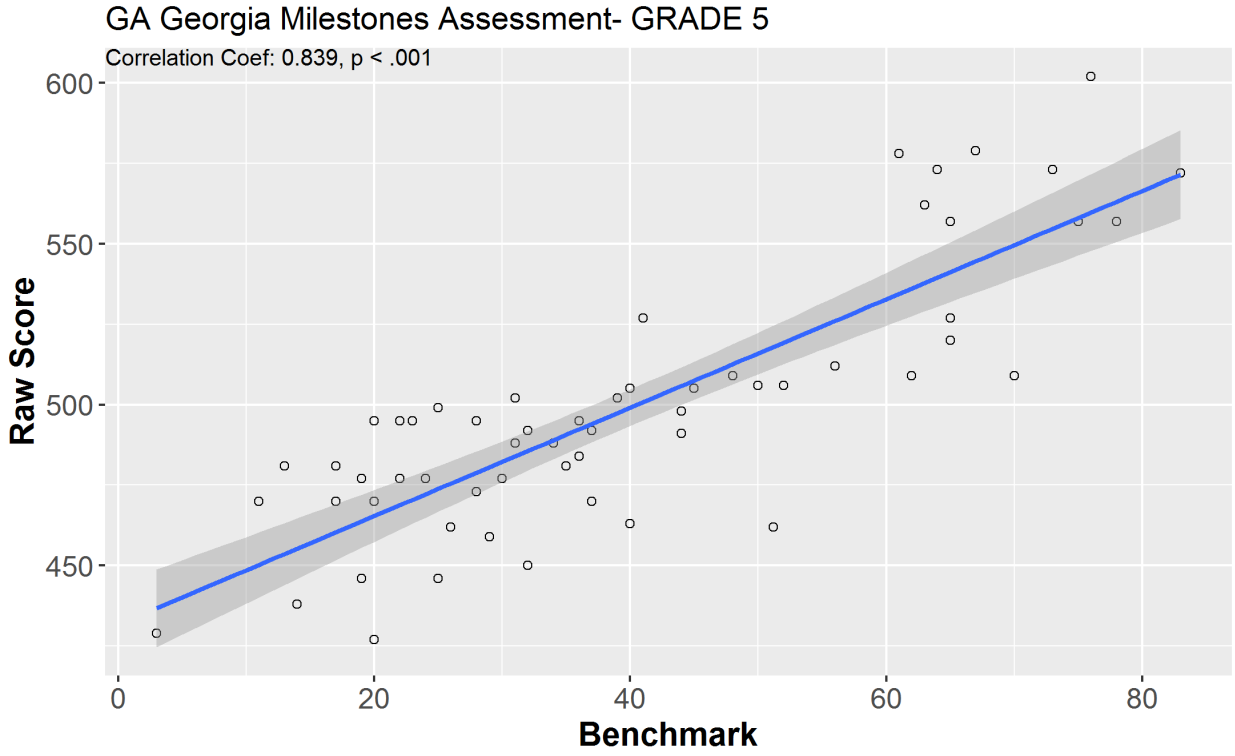

NWEA MAP- GRADE 3
Correlation Coef: 0.805, p < .001
Raw Score
220
210
200
190
180
20
40
60
80
Benchmark

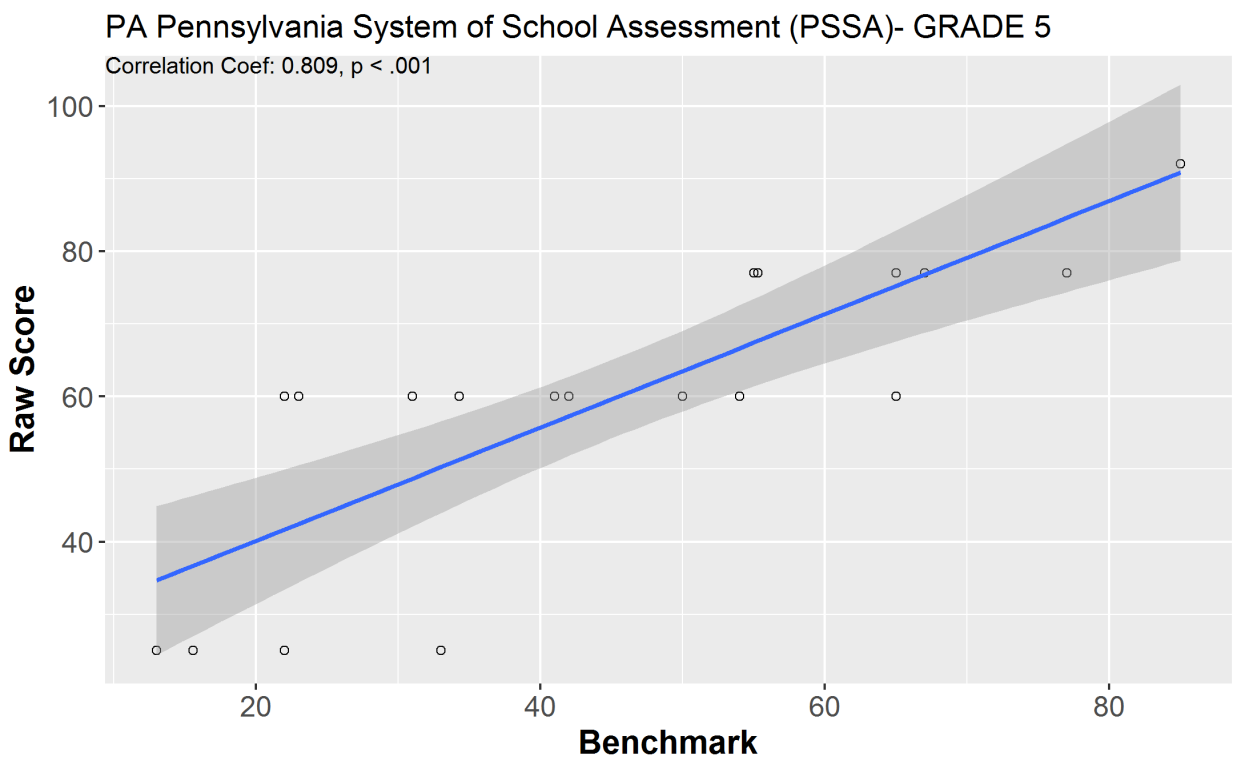

AL Scantron Performance Series- GRADE 3
Correlation Coef: 0.611, p = 0.001
Raw Score
2600
2400
2200
0
20
40
60
Benchmark

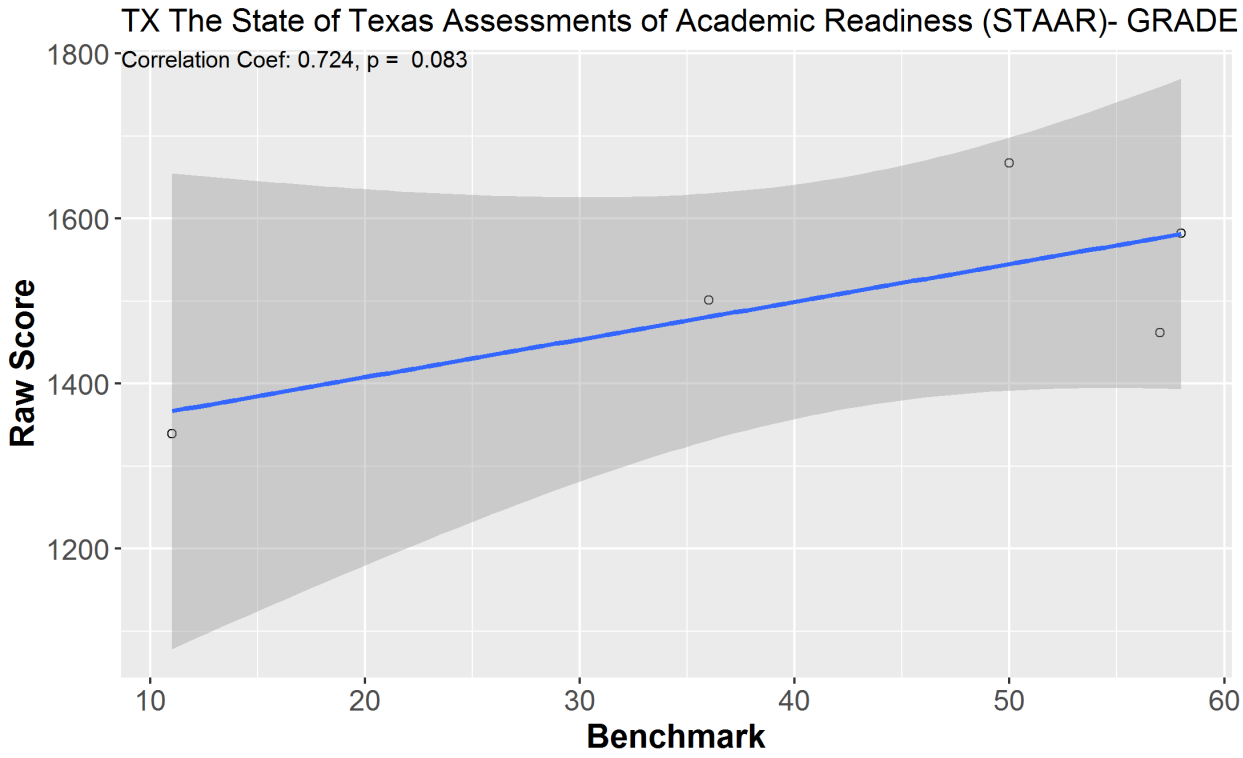

SD Smarter Balanced- GRADE 3
Correlation Coef: 0.784, p < .001
Raw Score
Benchmark
1
2
3
4
20
40
60
80


NC North Carolina End-of-Grade Test- GRADE 4
Correlation Coef: 0.733, p < .001
Raw Score
Benchmark
2
4
6
0
25
50
75
100


AL Scantron Performance Series- GRADE 5
Correlation Coef: 0.762, p < .001
Raw Score
2800
2600
2400
2200
20
40
60
80
Benchmark


GA Georgia Milestones Assessment- GRADE 5
Correlation Coef: 0.839, p < .001
Raw Score
600
550
500
450
0
20
40
60
80
Benchmark


PA Pennsylvania System of School Assessment (PSSA)- GRADE 5
Correlation Coef: 0.809, p < .001
Raw Score
Benchmark
100
80
60
40
20
40
60
80


TX The State of Texas Assessments of Academic Readiness (STAAR)- GRADE
Correlation Coef: 0.724, p = 0.083
Raw Score
Benchmark
1800
1600
1400
1200
10
20
30
40
50
60

DISCUSSION

The results of this correlation analysis indicate a strong relationship between MobyMax's Benchmarker Assessments for mathematics and state standardized tests in mathematics. No correlation coefficient in Table 1 is below .5. The fact that this relationship was shown over 5 grade levels and across seven states adds to the strength of the finding.

The strong relationship between the MobyMax's Benchmarker Assessments and the various state standardized test scores indicates that MobyMax's Benchmarker Assessments for mathematics have high validity. Students who score highly on MobyMax's Benchmarker Assessment are likely to score highly on the state standardized tests analyzed in this research and vice versa. In other words, there is strong evidence that MobyMax's Benchmark Assessments are measuring the same constructs that are being measured in the state standardized tests analyzed in this research.

This result indicates that MobyMax's Benchmarker Assessments for mathematics should be very useful for identifying learning gaps, determining which standards students need further instruction on, and helping teachers adjust instruction accordingly to better prepare their students for state standardized assessments.

REFERENCES

Chhabra, Geeta, Vasudha Vashist, Jayantyi Ranjan (2017). A Comparison of Multiple Imputation Methods for Data with Missing Values. Indian Journal of Science and Technology, Vol 10(19).

Cohen, Jacob (1988). Statistical power analysis for the behavioral sciences 2nd ed. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Greene, William H. (2002). Econometric Analysis 5th ed. Prentice Hall, Upper Saddle River, New Jersey.

Ozer, Daniel (1985). Correlation and the Coefficient of Determination, Psychological Bulletin · March 1985, pgs. 308-315