Comparability of MAP Growth Tests Administered through Different Technology and Psychometric Infrastructure: An Engine Evaluation Study based on Empirical Data

September 2020

Emily Bo, NWEA Psychometric Solutions
Patrick Meyer, NWEA Psychometric Solutions

nwea

© 2020 NWEA. NWEA and MAP Growth are registered trademarks of NWEA in the U.S. and in other countries. All rights reserved. No part of this document may be modified or further distributed without written permission from NWEA.

Suggested citation: Bo, E., & Meyer, P. (2020). *Comparability of MAP Growth tests administered through different technology and psychometric infrastructure: An engine evaluation study based on empirical data*. Portland, OR: NWEA.

Table of Contents

List of Tables

List of Figures

Executive Summary

NWEA® began an initiative dubbed "Project Altair" to enhance the constraint-based engine (CBE) originally designed for state summative assessments so it can also be used to deliver the MAP® Growth™ interim assessments. As part of the Project Altair initiative, new features were added to CBE between 2018 and 2019 to accommodate requirements for administering MAP Growth tests in a comparable way to the current MAP Growth engine known as COLO. After the CBE enhancement process was completed in December 2019, final simulations were conducted using simulated data on both CBE and COLO to check the mode comparability based on content validity, construct validity, test score reliability, adaptivity, and item exposure within and across administrations (Hu et al., 2020). This report presents results of an extension of the mode comparability study based on empirical data from the Fall 2019 and Winter 2020 MAP Growth results from Nebraska students.

In 2019, after the launch of Project Altair, NWEA implemented a school sampling procedure in Nebraska to ensure the rough equivalency of demographic characteristics between the COLO and CBE groups. In Winter 2020, some Nebraska schools took the MAP Growth tests on CBE and others took them on COLO. The study includes results from comparing student RIT scores, test length and duration, test score reliability, test content analysis, item exposure, and engine adaptivity. Below is a summary of the major findings and conclusions:

1. The differences in the MAP Growth Reading and Mathematics data are not practically significant in RIT scores, test length, and test duration between CBE and COLO. The tests delivered on the two engines have comparable test reliabilities that were all above 0.9. The CBE test scores had slightly higher precision than COLO's. Results confirmed the findings from the simulation study suggesting that the two engines are comparable in delivering valid and reliable MAP Growth tests (Hu et al., 2020).
2. Item exposure rates are comparable, with most items having an item exposure rate below 10%. The CBE Reading item-use rate is lower than COLO's. This difference was due to (1) the differences in the passage item selection algorithm and (2) the CBE sample size being much smaller than COLO's. Thus, there are more items than students at the lower end of the scale.
3. The content analysis shows that the tests delivered on CBE meet all the content constraints defined in the test models, including the total test length and number of operational and field test items, the passage-related constraints, and the instructional-level item count constraints. COLO, which primarily controls the instructional content areas, has wider item count ranges in the number of passages, number of items per passage, and number of items per instructional area. However, the overall results are comparable between the two engines.
4. CBE shows better engine adaptivity than COLO, especially for extremely low or high achievement students. The capability of CBE in providing higher score precision and better adaptivity is also confirmed by the Project Altair simulation study (Hu et al., 2020).

The focus of this study was to check that the two engines are comparable in delivering valid and reliable MAP Growth tests. Therefore, the CBE test models were specified to ensure minimum differences from the tests delivered on COLO. With almost equivalent specifications of the test blueprints, the results of this study indicate that the MAP Growth tests delivered on the two engines are comparable, although the CBE test scores are more precise than COLO's and CBE has better engine adaptivity than COLO.

1. Introduction

This document presents the results of a mode comparability study conducted by NWEA® based on empirical data to evaluate how scores from MAP® Growth™ administered on the constraint-based engine (CBE) compare to those administered on the current MAP Growth engine known as COLO. CBE was originally developed to deliver end-of-year state summative assessments, but it has been enhanced to also deliver interim tests such as MAP Growth multiple times throughout the year. While CBE and COLO differ in software infrastructure and psychometric adaptive algorithms, these differences should have little impact on test scores when delivering tests built from the same test blueprints. Results from this study will help determine if MAP Growth assessments can be administered on CBE in their current form without affecting scores.

1.1. Project Altair Overview

In 2018, NWEA developed a new adaptive testing engine known as CBE to support the state summative assessments. Given that CBE was originally designed for fixed-length adaptive tests administered to students once a year, an important goal of the Project Altair initiative was to enhance CBE so it can deliver variable-length interim assessments multiple times a year and produce comparable scores to COLO. New features were added to CBE to accommodate requirements for administering MAP Growth tests. Although efforts have been made to keep many of the features of CBE and COLO the same, there are fundamental differences between the item selection algorithms adopted for the two engines. Given the differences, research questions for the mode comparability study include the following:

1. Do CBE and COLO produce comparable scores for tests designed from the same test blueprints?
2. Do the two engines perform equally well in terms of adaptivity and item exposure control?

Two studies are being conducted as part of Project Altair to investigate mode comparability: (1) a study using simulated data and (2) a study based on empirical student data as presented in this report. The simulation study investigated whether the estimated scores are close to the true scores (Hu et al., 2020). Simulated data are not confounded by factors such as student motivation and unexpected issues that could happen during test administration and the data collection process. In the context of Project Altair, the simulation study could also help researchers discover any unexpected factors in the CBE enhancement process that could potentially cause mode incomparability before the empirical data are collected. This study complements the simulation results by using empirical data to determine the practical extent to which these differences have affected students' performance and their test-taking experience. To conduct the mode comparability study using actual student data, NWEA administered the same six MAP Growth Mathematics and Reading tests used in the simulations on both COLO and CBE to selected students in Nebraska in Winter 2019/2020.

1.2. CBE Enhancements

The MAP Growth tests delivered on CBE and the ones delivered on COLO share some common features, including the following:

1. Item pool, including the test items and item parameters
2. Student population
3. Online adaptive administration
4. Entry condition of a student's initial ability value
5. The use of maximum likelihood estimation (MLE) with fencing rules as the final ability estimation method
6. Test termination conditions

The following enhancements were added to CBE to improve the overall efficiency of the online test delivery and use of the MAP Growth item pools: (1) longitudinal item exposure COLO as a guideline rather than a constraint, (2) momentary ability estimation methods, and (3) item selection algorithm. The new enhancement features added to CBE may impact students' scores and their test-taking experience. For more information on these CBE enhancements, please refer to the Project Altair simulation report (Hu et al., 2020).

1.3. Literature Review

Most of the score comparability literature is about adaptive and paper-pencil testing (e.g., Pomplun & Custer, 2005; Tsai & Shin, 2013; Wang & Kolen, 2001). However, even though the literature does not provide the same mode comparison as the one in this study, literature reviews help to identify relevant methodologies and provide context for the results. Therefore, from the literature, there are three general categories of criteria adopted in evaluating comparability: (1) content and construct validity, (2) psychometric properties and reliability, and (3) statistical assumption/test administration condition that includes whether the assumptions used to establish comparability hold and whether the operational test conditions match the comparability study testing conditions

1.3.1. Content and Construct Validity

Content validity refers to the extent to which the items on a test are fairly representative of the entire domain the test seeks to measure. Construct validity is the degree to which a test measures what it claims to be measuring. The Project Altair simulation study showed that both COLO and CBE can recover the true abilities well (Hu et al., 2020), which provides evidence of the construct validity of the tests delivered on the two platforms. To support the content validity of MAP Growth, COLO and CBE are compared in terms of meeting the content specifications based on the number of items and passages from the empirical data.

1.3.2. Psychometric Properties and Reliability

The item selection algorithm involves three key components: content balancing, item selection criterion, and item exposure control. The check of content balancing is included as part of the content validity criterion. The item selection criterion can be assessed by comparing engine adaptivity between CBE and COLO or by comparing score precisions that reflect whether the engine selects items with maximized fisher information at a given ability value. Unlike linear tests with items designed for a single use during a test event, an adaptive test reuses all items in the item pool over time. Some items may be selected and used too frequently. Such excessive exposure of items to a test population could change a student's test-taking behavior regarding those compromised items, which could threaten the test's fairness and validity. Another aspect of the criterion is that items are not recalibrated after switching to CBE, so the psychometric property and reliability checkup will focus on the student Rasch Unit (RIT) scores. The criterion requires that the two score distributions have the same means and standard deviations.

1.3.3. Statistical Assumption/Test Administration Condition

Several studies have found differences between the scores for various subgroups (e.g., Eignor & Schaeffer, 1995; Segall, 1995). Thus, variables such as sex and race are included in this study to assess the relationship across modes. The test administration criterion involves the equivalence of the two modes regarding item presentation, data collection design, and the statistical analyses of the test scores. Most of the summarized scenarios in the literature are not applicable to this study. However, the criterion may be extended to the students' test-taking experience on the two engines. The difference of the item selection algorithm between COLO and CBE may or may not affect students' testing. The manifest variables to examine include test duration and test length. Analyzing test duration and test length may help to understand whether the students' test-taking experience is equivalent between COLO and CBE.

2. Method

A group of Nebraska schools took the MAP Growth Reading and Mathematics tests on CBE, while others remained on COLO. The schools were then matched on key variables. The CBE group was considered the treatment group, whereas the control group were schools that remained on COLO. The treatment effect is referred to as “Altair.” Several analyses were conducted to compare the performance of students in these two groups. The score comparability evaluation is an ongoing process, and comparisons will continue being made as more schools transition to CBE. When students transition to CBE will also be tracked, which will be used as a time-varying covariate to study the impact of the transition.

2.1. School Sampling Procedure

The school sampling procedure for this study used a combination of principal component analysis (PCA) and stratified random sampling to divide the schools into two groups (CBE vs. COLO). This procedure ensures a rough equivalency of each group’s joint distribution of the following demographic variables:

1. School size: The number of students in a school based on the test response file, student-level data file, and test-level data file
2. IEP: Individualized Education Program indicator (dichotomous variable)
3. FRL: Free-and-reduced lunch (dichotomous variable)
4. Disability: The original variable had 13 categories but was collapsed into two categories in the analysis (No Disability and Disability)
5. Performance Level: Below the Standards, Meets the Standards, and Exceeds the Standards
6. Sex: Female and male
7. Race: While the original variable had six categories, this analysis collapsed them into two (White and non-White)

The following steps were used to construct school-level factor scores to represent their socioeconomic and state performance composition of its student body, including the sex and race of its students:

Step 1: Transform the demographic variables by turning categorical variables into ratios (e.g., the ratio of IEP students in a school). The variable showing a dramatic skew (i.e., variable race) was log-transformed, and all variables were standardized.

Step 2: Apply PCA to the transformed variables to obtain the loadings of the first principal component. The loadings are multiplied by the transformed variables to obtain the school factor scores.

Step 3: Use the mean and standard deviation of the school factor scores to construct a normal distribution and stratify the values. Half the schools in each stratum are assigned to the CBE group, and the remaining schools are assigned to COLO.

To plan for the sample size, a power analysis was conducted to guide the sample size selection for the simple between- group comparisons. Table 2.1 presents the sample sizes needed to have an 80% chance of detecting an effect size of a given value. For example, to detect a difference of 1 RIT, the sample must include 9,076 students. Fewer students (N=1,010) are needed to detect a RIT difference of 3 points.

Table 2.1. Sample Size Needed to Detect Each Effect Size

#Students	Effect Size, *g*	RIT Score Difference
364	0.29	5
570	0.24	4
1,010	0.18	3
2,270	0.12	2
9,076	0.06	1

2.2. Descriptive Statistics

2.2.1. RIT Scores

The mean and standard deviation of the RIT scores were computed for each administration mode, grade, and term. The differences between scores were examined using Hedges' g effect size (Hedges, 1981), which is the standardized mean difference between the scores estimated by the two test administration modes. The effect size (g) is computed as follows:

$$g = (\bar{\theta}_x - \bar{\theta}_y)/S$$

where $\bar{\theta}_x$ and $\bar{\theta}_y$ are the mean RIT scores from the two test administration modes, and *S* is the pooled standard deviation calculated as follows:

$$S = \sqrt{\frac{(n_x - 1)(S_x)^2 + \left(n_y - 1\right)\left(S_y\right)^2}{n_x + n_y - 2}}$$

where n_x and n_y are the number of students who took the tests in each mode. The general guidelines of the reference values for the standardized effect size measures are 0.2, 0.5, and 0.8 corresponding to small, medium, and large effects.

2.2.2. Test Length and Duration

The differences of test length and test duration (in minutes) were examined across the two administration modes based on simple descriptive statistics and effect sizes with overall test duration and test length. MAP Growth tests delivered on COLO and CBE are both variable-length adaptive tests, so test length was determined by the number of items a student takes in a test. Test duration is the amount of time a student takes to finish the test. These variables are of interest as they reflect the actual interaction between students and the online test administration.

2.3. Marginal Reliability and SEM

To evaluate score reliability from the tests delivered on COLO and those delivered on CBE, marginal reliability and the mean standard error of measurement (SEM) were compared. Score precision of MAP Growth scores is measured by the SEM, a function of the relationship among item parameters, the ability of the student, and the number of items administered. SEM is related to reliability in that it estimates how repeated measures of a student on the same assessment tend to be distributed around their "true" score. The SEM is the inverse of the square root of test information. Score precision is best when students are given items closely matched to their abilities. Lower values of SEM indicate greater precision in the scores.

Marginal reliability (Samejima, 1977) measures how well the items on a test that reflect the same construct yield similar results. It is also an internal consistency measure. The approach taken for MAP Growth was suggested by Wright (1999) and is given by the following:

$$\rho_\theta = \frac{\sigma_\theta^2 - M_{S_\theta^2}}{\sigma_\theta^2}$$

where σ_θ^2 is the observed variance of the achievement estimates, θ and $M_{S_\theta^2}$ is the observed mean of the score's conditional error variances at each value of θ. Tests are considered of sound reliability when their marginal reliability coefficients range from 0.80 and above.

2.4. Mixed-Effect Model Fitting

Simple between-group comparisons provide a way to gauge the overall impact of moving from COLO to CBE, but they do not capture the effect on the school level. Because of the nested structure of the datasets (i.e., students are nested within schools), it is reasonable to assume that the school itself above and beyond the Altair treatment condition would have an impact on the performance of the students. This impact would manifest itself as correlations in achievement scores among students attending the same schools. A simple one-way comparison of the score means for the treatment and control groups would likely violate the assumption of independent errors because the school factor would have an additional impact on the test scores. Therefore, mixed-effect modeling was used to study the Altair effect in a nested structure, where students are nested within schools. Factors such as a student's grade, number of instructional days, sex, race, and past achievement scores would also affect their current RIT scores. For better evaluation, the goal was to isolate the Altair effect by including those factors into the model. The general model framework at each level is shown below:

Level 1 (Between Students / Within School):

$$Y_{ij} = \beta_{0j} + \boldsymbol{\beta Covariates_{ij}} + e_{ij},$$
$$e_{ij} \sim N(0, \sigma^2),$$

Level 2 (Between Schools):

$$\beta_{0j} = \gamma_{00} + \boldsymbol{b_0 ContextEffects_j} + \mu_{0j},$$
$$\mu_{0j} \sim N(0, \tau_{00}^2)$$

Intraclass correlation was used to estimate the correlation among individual students' scores within the nested structure (i.e., within schools). It is expressed as follows:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

where τ^2 is the population variance between clusters, and σ^2 is the population variance within clusters.

Higher values of ρ indicate that a greater share of the total variation in the outcome measure is associated with the cluster membership, meaning there is a relatively strong relationship among the scores of two students from the same cluster (school). To evaluate the models, the chi-square likelihood ratio test between successive models was conducted. One common situation of the successive models are the "nested" models, where one model is obtained from the other one by putting some of the parameters to be zero. The null hypothesis is "reduced model is true," and the alternative hypothesis is "current model is true." The likelihood ratio statistics is as follows:

$$\Delta G^2 = -2logL \text{ from reduced model} - (-2logL \text{ from current model})$$

and the degrees of freedom is *k* (the number of omitted predictors in the reduced model). ΔG^2 has a chi-square distribution and the p-value is $P(\chi_k^2 \geq \Delta G^2)$.

In statistical hypothesis testing, the *p*-value has been the standard approach to determine whether an estimated model coefficient is significant. However, *p*-values in mixed-effect modeling are not recommended because the distribution of the test statistics of the null hypothesis does not have a *t*-distribution. Therefore, an alternative approach is needed. This study used the profile confidence intervals to determine the significance of model predictor coefficients. An effect is not significant when 0 is in its profile confidence interval. The magnitude of an effect was evaluated using the effect size measure from Aiken and West (1991):

$$f^2 = \frac{R_2^2 - R_1^2}{1 - R_2^2}$$

where R_2^2 is the variance explained for a model with the given effect, and R_1^2 is the variance explained for a model without the given effect. The measure can be interpreted as the proportion of variance explained by the given effect relative to the proportion of outcome variance unexplained (Aiken & West, 1991) and is considered small at a value of 0.02, medium at a value of 0.15, and large at a value of 0.35 (Cohen, 1992).

2.5. Item Selection Process

Another essential part of the mode comparability evaluation was to compare the COLO and CBE item selection process, which is the core of an adaptive test, based on content differences, item exposure, and engine adaptivity. First, the content differences between the tests delivered on COLO and those on CBE were checked based on adherence to the blueprint,. A test blueprint specifies the content standards and enlists the skills and concepts that need to be tested for each content standard, along with the relative importance of each. On COLO, test blueprints were defined for item selection specifications to balance item counts at the instructional area level. On CBE, blueprint functionality was extended by defining test models in which both guidelines and constraints are used to optimize the content balance requirements and the maximization of ability score precisions. Guidelines are "nice-to-haves," and constraints are "must-haves." The enhanced functionality of CBE allows the minimum content balance requirements to be met and allows for increased precision and reliability of a student's score through the optimal use of an item pool and an optimal test design (e.g., by allowing flexible specifications of item positions and the use of friend and enemy items). COLO and CBE were compared based on how well they met the content specifications in the COLO blueprints and CBE test models in terms of test length, number of operational items, number of field test items, number of items per instructional area, number of passages, number of items per passage, and passage positions.

Second, the item exposure rate within an administration (e.g., Winter 2020) was calculated for each item in the item pool using the following formula:

$$Exposure\ Rate = \frac{Number\ of\ times\ an\ item\ appear\ in\ a\ test\ event}{Total\ number\ of\ test\ events\ in\ an\ administration}$$

Lastly, the engine adaptivity was evaluated via whether the difficulty of an item presented to a student matched the student's ability. MAP Growth is based on the Rasch model, so the following delta value was used to indicate the adaptivity:

$$Delta = Item\ Difficulty - Momentary\ Ability$$

3. Results

In both Reading and Mathematics, scores outside the range of 100 to 350 were excluded. When a student had more than one score in a term, the score with the lower SEM was retained. Data were from Grades K–8. No data in Grade 9 and above were collected on CBE. The original raw data have differences in the versions of MAP Growth tests taken on CBE and COLO. The version of the tests taken on both platforms were retained in the analysis. All the off-grade uses of the MAP Growth tests on COLO were also excluded, as the tests administered on CBE were only given to on-grade students.

3.1. Study Sample

The data include 171,093 students from 803 schools across 260 districts in Nebraska based on Fall 2019 school affiliations.[1] Table 3.1 presents the demographics of the study sample. In both content areas, more than half of students were White. The next highest race category was Hispanic, followed by Black, Asian/Pacific Islander (PI), and American Indian/Alaska Native (AI/AN). Males made up slightly more of the study sample than females.

Table 3.1. Study Sample Demographics

		%Students									
		Race*							Sex		
Grade	N	White	Hispanic	Black	Asian/PI	AI/AN	Other/MR	NA	Female	Male	NA
Reading											
K	13,911	59.22	21.52	9.27	4.23	0.91	4.85	–	48.06	51.41	0.53
1	14,177	59.57	21.57	8.90	4.03	0.93	5.00	–	48.24	51.72	0.04
2	17,181	61.15	20.10	8.72	3.91	1.07	5.06	–	49.22	50.74	0.03
3	20,707	60.55	18.98	7.29	3.46	1.03	8.60	0.09	48.96	51.04	–
4	21,280	60.24	19.39	7.54	3.22	1.13	8.46	0.02	48.36	51.63	0.01
5	21,711	60.01	19.93	7.35	3.14	1.11	8.43	0.02	48.89	51.05	0.06
6	20,268	60.20	18.84	7.71	3.09	1.25	8.88	0.02	48.51	51.47	0.02
7	18,290	58.88	19.64	7.87	2.99	1.49	9.11	0.02	48.70	51.26	0.03
8	18,551	60.69	19.82	7.88	3.15	1.40	7.05	0.02	48.89	51.01	0.10
Mathematics											
K	13,584	60.00	20.16	9.39	4.32	0.88	5.26	–	47.87	51.58	0.54
1	14,248	60.63	20.00	9.03	4.05	0.97	5.32	–	48.08	51.89	0.04
2	15,601	62.84	18.89	8.53	3.67	1.28	4.80	–	49.56	50.38	0.06
3	20,311	61.45	17.88	7.34	3.49	1.04	8.71	0.09	48.92	51.07	–
4	21,010	61.12	18.27	7.60	3.27	1.19	8.53	0.02	48.31	51.68	0.01
5	21,262	60.12	19.49	7.44	3.17	1.13	8.62	0.02	48.88	51.06	0.06
6	20,188	60.15	18.76	7.81	3.10	1.24	8.92	0.02	48.50	51.47	0.02
7	18,355	59.34	19.29	7.83	3.01	1.45	9.06	0.02	48.80	51.16	0.04
8	18,673	60.84	19.65	7.84	3.07	1.42	7.16	0.02	49.01	50.89	0.10

*Asian/PI = Asian/Pacific Islander. AI/AN = American Indian/Alaska Native. Other/MR = Other/Multi-Race.

[1] Winter 2020 had 751 school and 241 district affiliations.

3.2. Descriptive Statistics

3.2.1. RIT Scores

Figure 3.1 presents the trend of RIT mean scores by grade and term on both COLO and CBE. Blue represents COLO, and red represents CBE. The solid lines are winter scores, and the dash lines are fall scores. Table 3.2 presents RIT score descriptive statistics by content area, grade, and term for both CBE and COLO. As shown in the figure and table, COLO has higher RIT scores in most grades in both content areas and terms. Exceptions include the Reading RIT mean scores of both terms in Grades K–1, as the mean scores of CBE are higher than the mean scores of COLO. The mean scores increase as the grade level increases.

Figure 3.1. Mean RIT Scores by Grade

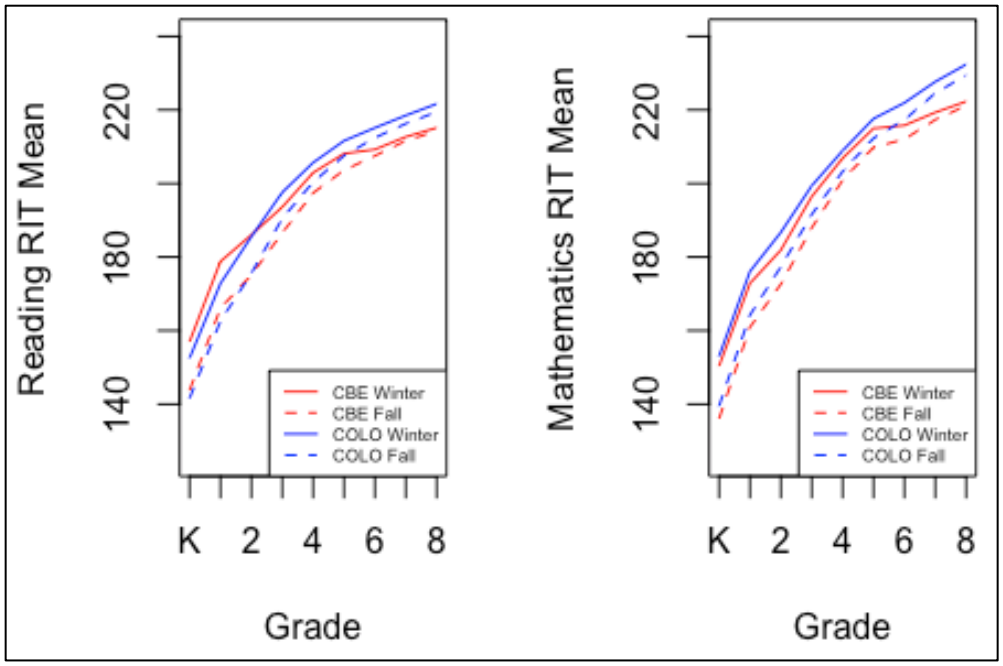


Table 3.2. Descriptive Statistics of MAP Growth RITs

Engine	Grade	#Students	Fall 2019		Winter 2020	
			Mean	SD	Mean	SD
Reading						
	Overall	166,076	194.58	27.74	200.30	25.05
CBE	K	64	143.74	8.50	156.96	11.02
	1	56	166.08	12.27	178.78	10.62
	2	40	175.11	15.73	186.02	14.60
	3	369	186.54	16.16	193.56	15.40
	4	950	197.41	15.71	202.92	15.07
	5	777	203.59	15.27	208.11	14.20
	6	1,077	207.51	16.72	209.24	17.23
	7	770	211.61	16.94	212.67	17.77
	8	823	214.23	16.56	215.13	17.94
	CBE Total	4,926	203.58	19.76	206.97	18.57

Engine	Grade	#Students	Fall 2019		Winter 2020	
			Mean	SD	Mean	SD
COLO	K	13,847	141.55	9.91	152.48	11.75
	1	14,121	162.57	13.77	172.79	14.44
	2	17,141	175.53	16.16	185.52	15.79
	3	20,338	190.47	15.81	197.57	15.18
	4	20,330	200.37	15.72	205.66	14.88
	5	20,934	207.51	15.27	211.60	14.53
	6	19,191	212.43	14.90	215.12	14.47
	7	17,520	216.26	15.33	218.49	15.01
	8	17,728	219.45	15.31	221.63	14.95
	COLO Total	161,150	194.30	27.91	200.10	25.19
Mathematics						
	Overall	163,232	198.70	30.19	205.39	27.49
CBE	K	555	136.04	11.52	150.34	13.63
	1	485	160.91	14.76	172.83	14.51
	2	634	172.60	12.60	181.89	12.61
	3	1,024	187.91	13.42	196.41	13.78
	4	1,171	200.90	14.03	206.87	14.62
	5	969	209.81	15.49	214.95	16.04
	6	1,121	212.03	15.95	215.80	17.06
	7	798	217.32	17.50	219.31	18.35
	8	852	221.15	17.98	222.29	19.08
	CBE Total	7,609	196.28	28.41	202.46	25.83
COLO	K	13,029	139.45	12.47	153.00	13.73
	1	13,763	164.21	14.63	176.06	14.25
	2	14,967	177.35	13.31	186.71	12.80
	3	19,287	191.54	13.03	199.36	12.83
	4	19,839	203.25	13.96	208.90	13.73
	5	20,293	212.30	15.15	217.61	15.63
	6	19,067	217.28	14.50	221.92	15.13
	7	17,557	224.54	16.18	227.66	16.66
	8	17,821	229.44	17.36	232.36	17.88
	COLO Total	155,623	198.82	30.27	205.53	27.56

To account for the students' growth between Fall 2019 and Winter 2020, the variable "Time" indicates the number of days between the fall and winter test dates. The variable is a proxy of the instructional days. In Reading, the mean of "Time" is 117.49 and its standard deviation is 20.27. The gap between the test dates ranges between 28 to 201 days. In Mathematics, "Time" has a mean of 121.24 and a standard deviation of 17.89 and ranges from 24 to 204 days. The distribution of the "Time" variable is roughly normal for both content area as their skewness and kurtosis are all near 0.0. The variable is later used in the mixed-effect modeling.

Table 3.3 presents the Hedge's g results by grade and term. The values in the last column are the effect size measure of growth. The growth scores were calculated by subtracting the fall RIT scores from the winter RIT scores. The growth score accounts for students' academic differences before the winter and is therefore a better measure to quantify the Altair effect. A positive effect size value ($g > 0$) of a growth score indicates that the CBE group has a higher growth score than the COLO group, whereas a negative effect size value ($g < 0$) indicates that the COLO group has a higher growth score. As shown in the last column, only Grades K–1 have the absolute effect size values above 0.2 in Reading. In Mathematics, only Grade 8 has the absolute effect size value above 0.2. In all three cases (i.e., Reading Grade K, Reading Grade 1, and Mathematics Grade 8), the Altair effects are small according to Cohen's guidelines. Among the three cases, the CBE groups in Grades K and 1 have higher growth scores in Reading than the COLO group. The opposite is observed in the Mathematics Grade 8 growth scores.

Table 3.3. Hedge's *g* Effect Size of MAP Growth RITs

Grade	Hedge's *g* Effect Size		
	Fall 2019	Winter 2020	Growth (Winter – Fall)
Reading			
K	0.22	0.38	0.28
1	0.26	0.42	0.32
2	-0.03	0.03	0.11
3	-0.25	-0.26	-0.01
4	-0.19	-0.18	0.03
5	-0.26	-0.24	0.06
6	-0.33	-0.40	-0.13
7	-0.30	-0.38	-0.15
8	-0.34	-0.43	-0.16
Mathematics			
K	-0.28	-0.19	0.09
1	-0.23	-0.23	0.01
2	-0.36	-0.38	-0.01
3	-0.28	-0.23	0.11
4	-0.17	-0.15	0.05
5	-0.16	-0.17	-0.03
6	-0.36	-0.40	-0.14
7	-0.44	-0.50	-0.18
8	-0.48	-0.56	-0.28

3.2.2. Test Length and Duration

Table 3.4 presents the mean and standard deviation (SD) of test length and duration (in minutes) by content area and engine. CBE and COLO have comparable test lengths averaging around 40–41 items per Reading test and 50–51 items per Mathematics test. Notable differences are observed for test duration as the CBE group has higher test durations than the COLO group in both terms and content areas, with the differences being more prominent in Reading. Similar magnitude and direction of the differences are observed in both terms, indicating that the differences might be caused by factors other than the Altair effect.

Table 3.4. Descriptive Statistics of Test length and Test Duration

		Test Length				Test Duration			
		Fall 2019		Winter 2020		Fall 2019		Winter 2020	
Engine	Count	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Reading									
COLO	161,150	40.70	1.24	40.68	1.23	56.37	29.17	59.75	30.18
CBE	4,926	40.28	0.81	41.02	0.61	67.52	32.03	70.24	34.13
Total	166,076	40.69	1.23	40.69	1.21	56.70	29.32	60.06	30.36
Mathematics									
COLO	155,623	50.68	3.68	50.78	3.70	54.47	26.80	57.99	28.29
CBE	7,609	51.06	3.32	50.55	3.14	56.09	29.17	59.21	29.60
Total	163,232	50.70	3.66	50.77	3.68	54.55	26.92	58.05	28.35

To further check the treatment effect in the test length and test duration data, effect size measures were calculated and are presented in Table 3.5. A positive effect size value ($g > 0$) indicates that the CBE group has a higher difference between winter and fall in test length and duration than the COLO group, whereas a negative effect size value ($g < 0$) indicates that the COLO group has a higher difference between winter and fall in test length and duration. As shown in the table, medium to large effect sizes (above 0.5) are observed in the differences of test length in most grades. However, the calculation indicates the number of items. An effect size of 1 is usually considered quite large, but in this application it indicates a difference of one item. Small effect sizes (between 0.2 to 0.3) are observed in Grades 1, 5, and 7 in the test duration differences of the Reading tests and in Grade K in the test duration differences of the Mathematics tests. Among these four cases, only the test duration difference in Reading Grade 7 has a negative value in its effect size value, indicating that the COLO group has a higher difference between winter and fall in test duration than the CBE group.

Table 3.5. Effect Size Results of Test Length and Test Duration

	Hedge's *g* Effect Size					
	Test Length			Test Duration		
Grade	Fall 2019	Winter 2020	Difference (Winter- Fall)	Fall 2019	Winter 2020	Difference (Winter- Fall)
Reading						
K	–	–	–	0.09	0.07	-0.02
1	–	0.01	0.01	0.23	0.43	0.25
2	0.12	1.15	0.69	0.24	0.37	0.16
3	-0.03	1.48	1.06	-0.17	-0.25	-0.10
4	0.02	1.45	1.03	-0.08	-0.06	0.02
5	0.01	0.12	0.09	-0.20	0.01	0.23
6	0.09	1.76	1.21	-0.01	0.01	0.02
7	0.05	1.45	1.08	0.41	0.11	-0.35
8	0.00	1.30	1.04	0.48	0.42	-0.01

Grade	Hedge's *g* Effect Size					
	Test Length			Test Duration		
	Fall 2019	Winter 2020	Difference (Winter- Fall)	Fall 2019	Winter 2020	Difference (Winter- Fall)
Mathematics						
K	–	–	–	-0.20	0.10	0.25
1	–	-0.01	-0.01	-0.14	-0.15	-0.02
2	0.25	-0.08	-0.55	-0.06	-0.03	0.03
3	-0.07	-1.01	-0.64	-0.07	-0.14	-0.08
4	-0.06	-0.97	-0.62	-0.04	-0.05	-0.02
5	-0.10	-1.13	-0.69	-0.03	-0.05	-0.03
6	0.09	-0.87	-0.67	0.01	-0.04	-0.05
7	0.01	-0.89	-0.63	0.19	0.07	-0.15
8	-0.01	-0.86	-0.59	0.22	0.28	0.10

3.3. Marginal Reliability and SEM

Table 3.6 presents the marginal reliabilities and SEM of winter RIT scores by content area, engine, and grade. The marginal reliabilities for all grades and both engines are in the 0.90s, which suggests that MAP Growth tests on both CBE and COLO have high internal consistency. The reliabilities are comparable across the two groups, with the largest difference observed in Reading Grade 1 with a value of 0.04. In general, the tests delivered on CBE have slightly higher precision than those delivered on COLO.

Table 3.6. Marginal Reliability and Mean SEM of Winter RIT Scores

Grade	N-Count		Mean SEM		Reliability	
	COLO	CBE	COLO	CBE	COLO	CBE
Reading						
K	13,847	64	3.24	3.18	0.92	0.92
1	14,121	56	3.23	3.18	0.95	0.91
2	17,141	40	3.37	3.27	0.95	0.95
3	20,338	369	3.36	3.28	0.95	0.95
4	20,330	950	3.36	3.27	0.95	0.95
5	20,934	777	3.37	3.29	0.95	0.95
6	19,191	1,077	3.35	3.28	0.95	0.96
7	17,520	770	3.36	3.28	0.95	0.97
8	17,728	823	3.36	3.28	0.95	0.97
Mathematics						
K	13,029	555	3.25	3.24	0.94	0.94
1	13,763	485	3.26	3.23	0.95	0.95
2	14,967	634	2.95	2.92	0.95	0.95
3	19,287	1,024	2.93	2.91	0.95	0.96
4	19,839	1,171	2.98	2.95	0.95	0.96
5	20,293	969	3.04	2.98	0.96	0.97
6	19,067	1,121	2.91	2.90	0.96	0.97
7	17,557	798	2.91	2.90	0.97	0.98
8	17,821	852	2.92	2.90	0.97	0.98

3.4. Mixed-Effect Model Fitting

This section presents results of the analysis using multilevel mixed-effect models. R (R Core Team, 2018) with lme4 (Bates et al., 2015) was used to perform the modeling analysis.

3.4.1. RIT Scores

3.4.1.1. Models

Table 3.7 presents the models used for data analysis. The corresponding R code is presented in Appendix A. The base models for RIT scores are as follows:

- *Model 1: One-way random effects ANOVA.* In this unconditional model, school is a random effect. Students were nested within schools. There is not an independent variable, but only the intercept. This model is useful for obtaining estimates of the residual and intercept variance when only the clustering by school is considered. The model is especially helpful for understanding the structure of the data, and the estimates of the variance among the students σ^2 and among the schools τ^2 can be used to estimate ρ (the intraclass correlation). Using the values presented in Table 3.8, the values of ρ for both content areas can be calculated. In Reading, the value would be $\rho = 0.34$, which indicates that the correlation of the winter Reading RIT scores among students within the same schools is approximately 0.34. In Mathematics, the intraclass correlation is $\rho = 0.41$, indicating the correlation of the winter Mathematics RIT scores among students within the same schools is approximately 0.41.

- *Model 2: Model with covariates except the Altair treatment effect.* Model 2 introduces three covariates to account for variability in students' winter Reading RIT scores: students' fall Reading RIT scores ("FallRit"), number of days between the two test dates ("Time"), and students' grade (referred as "Grade"). Sex and race are also included for context on the effect of variables for each school. The original race variable is recoded into two groups only (White and Other). The recoded race variable was recoded as "White." The sex variable coded 0 as *female* and 1 as *male* and was renamed as "Male." The percentage of male students ("MalePct") and the percentage of white students ("WhitePct") were calculated for each school as context effects in the model to adjust for the percentage of different sex and race groups in each school. All the variables were grand-mean centered before being included in the model, and the grade variable was adjusted by resetting the base grade being Grade 3. After excluding the cases that did not have either sex or race information, the data included a total of 165,906 students and 792 schools for Reading and 163,058 students and 788 schools for Mathematics.

  Winter Reading RIT scores were predicted from the grand mean-centered fall Reading RIT scores, the grand mean-centered time variable, the based-grade adjusted grade variable, and the sex and race variables by allowing effects of demographic variables (percentage of male and percentage of white of each school) to vary from one school to another. In other words, Model 2 has five fixed effect predictors (FallRit, Time, Grade, Male, and White), two random slopes in a school (WhitePct and MalePct), and a school-specific random intercept. The two random slopes were treated as correlated with one another. The estimates of random coefficient terms at the school level show that the sex-related variable explains more variability in winter RIT scores across schools than the race-related variable across schools in both content areas.

The models with Altair effects are as follows:

- *Model 3: Model with the Altair treatment effect.* This model examines the Altair effect. A dichotomous Altair variable was added as a fixed effect predictor into Model 2. The chi-square test for deviance of Model 2 and Model 3 yielded nonsignificant chi-squares for both content areas (chi-squared = 0.04, p= 0.83 in Reading; chi-squared = 1.06, p= 0.30 in Mathematics), indicating that the Altair variable is not needed to predict the winter RIT scores in both content areas.

- *Model 4: Model with Altair by sex interaction effect.* This model includes the interaction effect of Altair by sex to Model 3. The chi-square test for deviance of Model 3 and Model 4 yields a significant chi-square in Reading (chi-squared = 6.75, $p < 0.01$) and a non-significant chi-square in Mathematics (chi-squared=0.31, p=0.57), indicating that the sex and Altair interaction variable should be included to predict the winter Reading RIT scores.

- *Model 5: Model with both Altair by race and Altair by sex interaction effects.* This model is the result of Altair by race being added to Model 4. The chi-square test for deviance of Model 4 and Model 5 yields a non-significant chi-square in both content areas (chi-squared = 0.05, p= 0.82 in Reading; chi-squared = 0.62, p= 0.43 in Mathematics), indicating that the race and Altair interaction variable is not significant in predicting the winter RIT scores in both content areas.

Table 3.7. Mixed-Effect Models used for Data Analysis—RIT Scores

Model 1	Level 1: $Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 2	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{FallRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 3	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{FallRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + \beta_6 \mathrm{Altair_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 4	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{FallRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + \beta_6 \mathrm{Altair_{ij}} + \beta_7 \mathrm{Altair_{ij}} * \mathrm{Male_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$

Model 5	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \text{FallRit}_{\text{ij}} + \beta_2 \text{Time}_{\text{ij}} + \beta_3 \text{Grade}_{\text{ij}} + \beta_4 \text{White}_{\text{ij}} + \beta_5 \text{Male}_{\text{ij}} + \beta_6 \text{Altair}_{\text{ij}} + \beta_7 \text{Altair}_{\text{ij}} * \text{Male}_{\text{ij}} + \beta_8 \text{Altair}_{\text{ij}} * \text{White}_{\text{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \text{MalePct}_{\text{j}} + b_{1j} \text{WhitePct}_{\text{j}} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$

Table 3.8. Random Effects—RIT Scores

Model	Group	Name	Variance	SD	Correlation	
					(Intercept)	MalePct
Reading						
Model 1	School	(Intercept)	213.00	14.59	–	–
	Residual	–	411.10	20.28	–	–
Model 2	School	(Intercept)	2.53	1.59	–	–
		MalePct	8.01	2.83	0.47	–
		WhitePct	0.36	0.60	0.40	0.11
	Residual	–	56.74	7.53	–	–
Model 3	School	(Intercept)	2.53	1.59	–	–
		MalePct	8.00	2.83	0.47	–
		WhitePct	0.36	0.60	0.40	0.11
	Residual	–	56.74	7.53	–	–
Model 4	School	(Intercept)	2.53	1.59	–	–
		MalePct	7.78	2.79	0.45	–
		WhitePct	0.36	0.60	0.40	0.06
	Residual	–	56.74	7.53	–	–
Model 5	School	(Intercept)	2.53	1.59	–	–
		MalePct	7.78	2.79	0.45	–
		WhitePct	0.36	0.60	0.39	0.06
	Residual	–	56.74	7.53	–	–
Mathematics						
Model 1	School	(Intercept)	313.20	17.70	–	–
	Residual	–	452.10	21.26	–	–
Model 2	School	(Intercept)	1.71	1.31	–	–
		MalePct	7.53	2.74	0.51	–
		WhitePct	0.48	0.70	0.06	0.38
	Residual	–	41.35	6.43	–	–

Model	Group	Name	Variance	SD	Correlation	
					(Intercept)	MalePct
Model 3	School	(Intercept)	1.70	1.31	–	–
		MalePct	7.53	2.75	0.51	–
		WhitePct	0.49	0.70	0.06	0.37
	Residual	–	41.35	6.43	–	–
Model 4	School	(Intercept)	1.70	1.30	–	–
		MalePct	7.34	2.71	0.52	–
		WhitePct	0.49	0.70	0.06	0.38
	Residual	–	41.35	6.43	–	–
Model 5	School	(Intercept)	1.70	1.30	–	–
		MalePct	7.36	2.71	0.52	–
		WhitePct	0.48	0.69	0.06	0.40
	Residual	–	41.35	6.43	–	–

3.4.1.2. Results

Table 3.9 summarizes the fixed effect coefficients for all models in both content areas. In Reading, the estimated intercept of Model 2 is 199.34, which is a Grade 3 white male student's winter Reading RIT score adjusted for their fall RIT score, the instructional days, and the school differences. The coefficient of race variable is 0.94, meaning that white students are on average 0.94 higher than the other race students in winter Reading RIT scores. In Mathematics, the estimated intercept is 205.06 in Model 2, which is a Grade 3 white male student's winter RIT score adjusted for their fall RIT score, the instructional days, and the school differences. The coefficient of race variable is 0.76, meaning that white students are on average 0.76 higher than the other race students in winter Mathematics RIT scores.

Table 3.9. Fixed Effects—RIT Scores

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Reading										
(Intercept)	200.73	0.54	199.34	0.08	199.29	0.21	199.58	0.24	199.55	0.26
FallRit	–	–	0.83	0.00	0.83	0.00	0.83	0.00	0.83	0.00
Time	–	–	0.02	0.00	0.02	0.00	0.02	0.00	0.02	0.00
Grade	–	–	0.49	0.02	0.49	0.02	0.49	0.02	0.49	0.02
Altair	–	–	–	–	0.04	0.20	-0.25	0.23	-0.23	0.26
White	–	–	0.94	0.05	0.94	0.05	0.94	0.05	1.00	0.27
Male	–	–	-0.20	0.04	-0.20	0.04	-0.76	0.22	-0.76	0.22
Altair:White	–	–	–	–	–	–	–	–	-0.06	0.27
Altair:Male	–	–	–	–	–	–	0.58	0.22	0.58	0.22

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Mathematics										
(Intercept)	206.06	0.65	205.06	0.07	204.90	0.18	204.94	0.19	204.86	0.21
FallRit	–	–	0.89	0.00	0.89	0.00	0.89	0.00	0.89	0.00
Time	–	–	0.03	0.00	0.03	0.00	0.03	0.00	0.03	0.00
Grade	–	–	-0.23	0.02	-0.23	0.02	-0.23	0.02	-0.23	0.02
Altair	–	–	–	–	0.18	0.17	0.13	0.19	0.21	0.21
White	–	–	0.76	0.04	0.76	0.04	0.76	0.04	0.91	0.19
Male	–	–	0.44	0.03	0.44	0.03	0.36	0.15	0.36	0.15
Altair:White	–	–	–	–	–	–	–	–	-0.15	0.19
Altair:Male	–	–	–	–	–	–	0.09	0.15	0.09	0.15

Table 3.10 presents the model comparison results on the adjacent models for both content areas. The results show that after adding an interaction variable of gender and Altair to Model 3, Model 4 significantly improves the fit of the model for Reading. For Mathematics, the chi-square tests show that the added variables in Models 3, 4 and 5 do not improve the fit of the model. Thus, based on these model comparison results, Model 4 with the Altair and the Altair by sex interaction variables is the most promising model for Reading, and Model 2 without any Altair effect variable is the best choice for Mathematics.

Table 3.10. Mixed-Effect Model Comparisons

Models	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr(>Chisq)
Reading								
Model 2	13	1,142,444	1,142,575	-571,209	1,142,418	0.04	1	0.84
Model 3	14	1,142,446	1,142,586	-571,209	1,142,418			
Model 3	14	1,142,446	1,142,586	-571,209	1,142,418	6.75	1	0.01
Model 4	15	1,142,441	1,142,592	-571,206	1,142,411			
Model 4	15	1,142,441	1,142,592	-571,206	1,142,411	0.05	1	0.82
Model 5	16	1,142,443	1,142,604	-571,206	1,142,411			
Mathematics								
Model 2	13	1,071,239	1,071,369	-5.36E+05	1,071,213	1.06	1	0.30
Model 3	14	1,071,240	1,071,380	-5.36E+05	1,071,212			
Model 3	14	1,071,240	1,071,380	-5.36E+05	1,071,212	0.31	1	0.58
Model 4	15	1,071,242	1,071,392	-5.36E+05	1,071,212			
Model 4	15	1,071,242	1,071,392	-5.36E+05	1,071,212	0.62	1	0.43
Model 5	16	1,071,243	1,071,403	-5.36E+05	1,071,211			

The profile confidence intervals of all the coefficients were calculated for Reading in Model 4 and for Mathematics for Model 3. Table 3.11 presents profile confidence interval results. All the fixed effect variables other than the Altair variable are significant. The Altair variable coefficient has 0 in both the 95% confidence intervals of (-0.70, 0.20) in Reading and of (-0.16, 0.51) in Mathematics. The Altair by sex interaction coefficient is significant in Reading. Further, the correlation between the two context effect variables are not significant in both content areas.

Table 3.11. Profile Confidence Interval Results of Model Coefficients

	2.50%	97.50%
Reading Model 4		
cor_WhitePct.MalePct|SCHOOL	-0.71	0.78
(Intercept)	199.11	200.04
FallRit	0.83	0.83
Time	0.01	0.02
Grade	0.45	0.52
White	0.84	1.05
Altair	-0.70	0.20
Male	-1.19	-0.33
Altair:Male	0.14	1.02
Mathematics Model 3		
cor_WhitePct.MalePct|SCHOOL	-0.85	1.00
(Intercept)	204.55	205.24
FallRit	0.89	0.90
Time	0.03	0.04
Grade	-0.26	-0.19
White	0.67	0.85
Male	0.38	0.51
Altair	-0.16	0.51

Table 3.12 presents the following:

- RB1 and RB2: Explained variance at Level 1 and Level 2 (Raudenbush & Bryk, 2002, pp. 74 and 79)
- SB: Total variance explained according to Snijders and Bosker (2012, p. 112)
- MVP: Total variance explained based on "multilevel variance partitioning" as proposed by LaHuis et al. (2014)

RB1 and RB2 have been criticized for the potential to yield negative estimates when implemented in a sample (Hox, 2010; Jaeger et al., 2017; Kreft & de Leeuw, 1998; LaHuis et al., 2014; McCoach & Black, 2008; Nakagawa & Schielzeth, 2013; Recchia, 2010; Roberts et al., 2011; Wang et al., 2011). The measure suggested by Snijders and Bosker (2012) is perhaps the most widely used, so this study used the SB estimator in the subsequent analysis. The *mitml* package (Grund et al., 2016) in R was used to calculate them. The results in Table 3.12 show that the chosen models of both content areas explain above 90% of the variances in the data. For both content areas, the added Altair effect does not make any difference even in the third decimal in the explained variance. As the numerator of the f^2 is the difference between the corresponding models' variance explained, the f^2 values of Model 3 & Model 2 and Model 3 & Model 4 in Reading are both around 0. The f^2 value of Model 3 & Model 2 in Mathematics is also around 0. Thus, even though some of the Altair-related variables are statistically significant in the datasets, they are not practically significant as they contribute minimally in explaining the variances in the RIT score data. Furthermore, the Altair and the Altair-related interaction effects are negligible in both content areas.

Table 3.12. Results of the Variance Explained

Model	RB1	RB2	SB	MVP
Reading				
Model 2	0.862	0.988	0.905	0.907
Model 3	0.862	0.988	0.905	0.907
Model 4	0.862	0.988	0.905	0.907
Mathematics				
Model 2	0.909	0.995	0.944	0.942
Model 3	0.909	0.995	0.944	0.942

Both the model comparison indexes and the effect size measure results show that the Altair and Altair interaction variables do not contribute significantly in predicting the winter RIT scores. Thus, Model 2 is the best-fitting model for both content areas. Figure 3.2 plots the predicted mean RITs based on Model 2 and Model 5. The two sets of dots overlap with each other, indicating that the added Altair-related variables do not contribute meaningfully to the predicted RIT scores. Figure 3.3 is a plot of the observed and the predicted winter mean RIT scores based on Model 2 by grade. The observed winter RIT mean scores are in blue, and the predicted values are in red. The two sets are very close to each other, indicating that the predictions based on Model 2 are close to the observed values.

Figure 3.2. Predicted RIT Mean Scores for Model 2 vs. Model 5

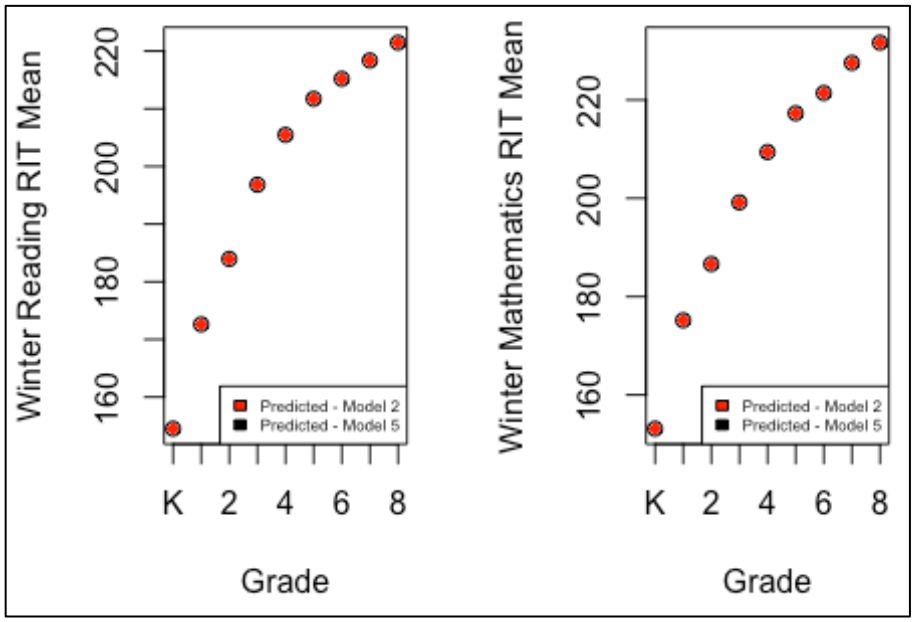
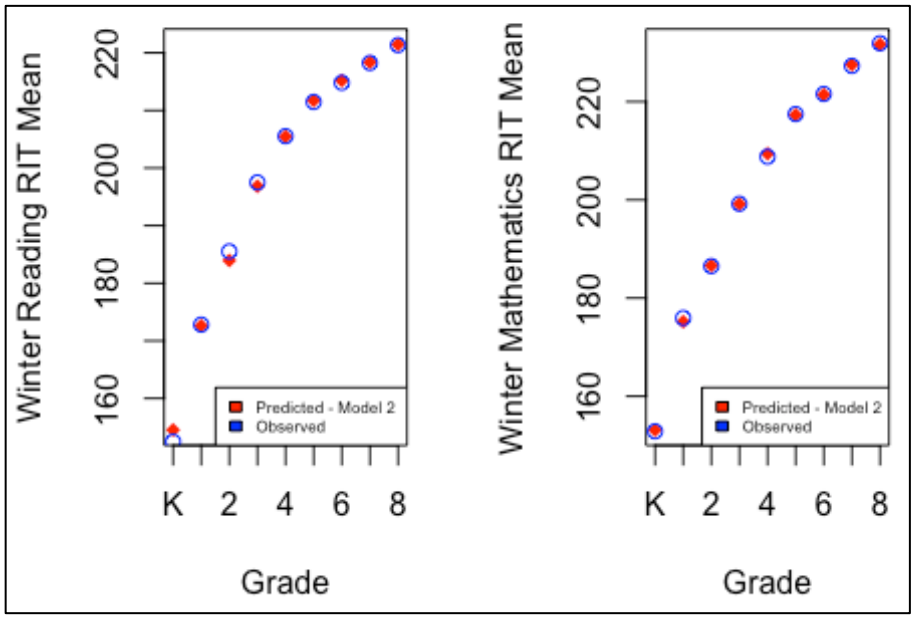


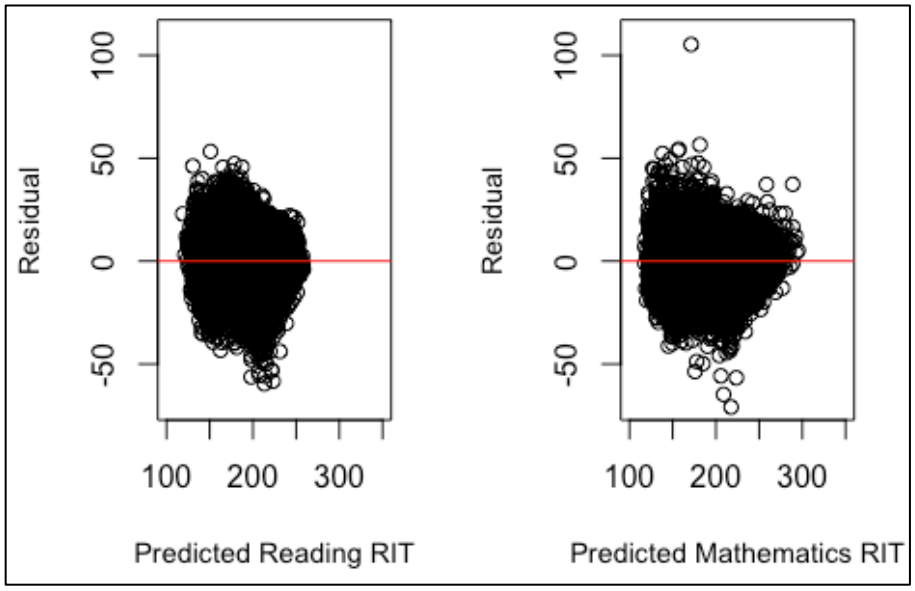
Figure 3.3. Observed vs. Predicted RIT Mean Scores



3.4.1.3. Residual Checks

With Model 2 being the chosen model to predict winter RIT scores, the model assumption of the residuals was further checked. The residuals are assumed to have a constant variance, be approximately normally distributed (with a mean of zero), and be independent of one other. One way to analyze residuals is a residual by predicted plot, which is a graph of each residual value plotted against the corresponding predicted value. If the assumptions are met, the residuals will randomly scatter around the center line of zero, with no obvious pattern. As shown in Figure 3.4, the residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable. The residuals roughly form a "horizontal band" around the 0 line, which suggests that the variances of the error terms are equal. No one residual "stands out" from the basic random pattern of residuals, which suggests that there are no outliers.

Figure 3.4. Residual by Predicted Plots



Next, whether the residuals of the model are normally distributed (at both levels) was checked. In addition to residuals being normally distributed, a multilevel model assumes that variance of the residuals is equal across groups (schools) for the different random effects. The Q-Q plots in Figure 3.5 and Figure 3.6 indicate that the assumptions might be violated at both levels in the two models of both content areas.

Figure 3.5. Residual Q-Q Plots—Reading

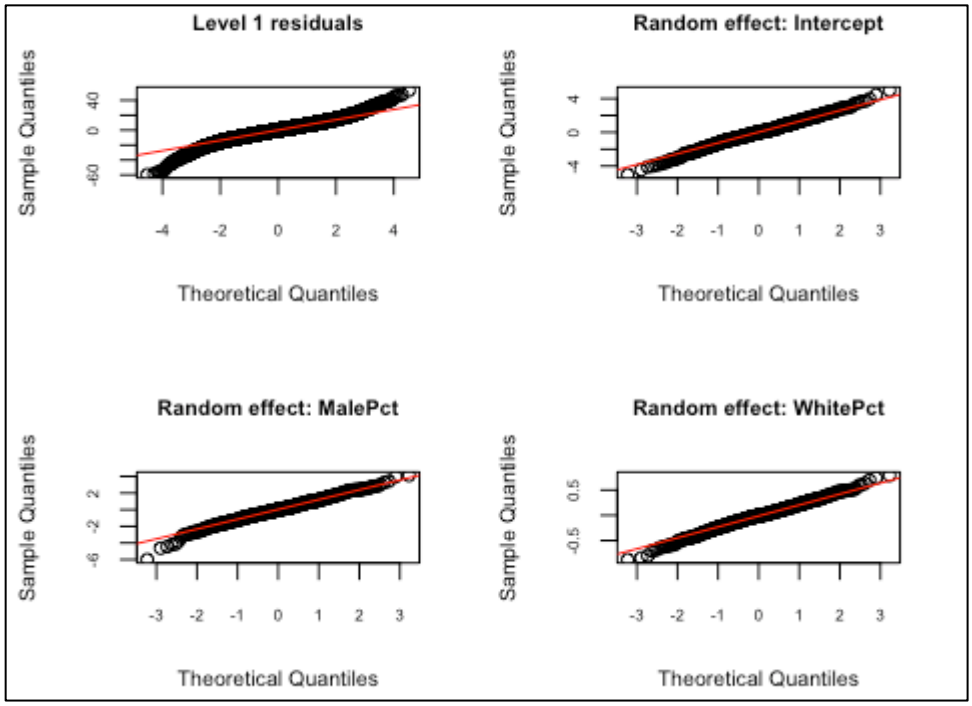
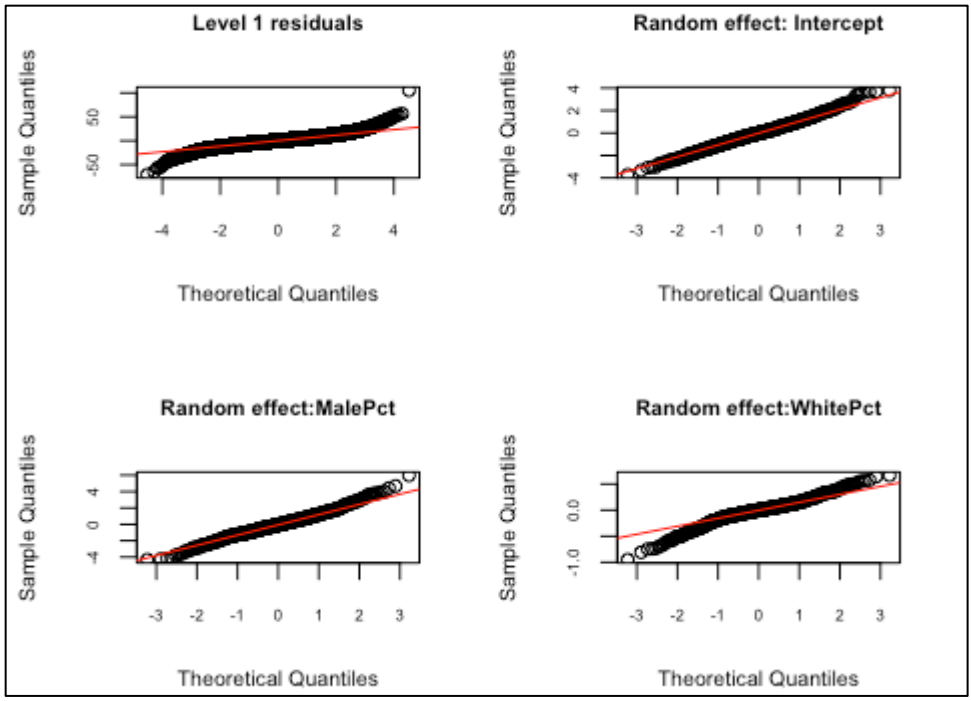


Figure 3.6. Residual Q-Q Plots—Mathematics



3.4.2. Test Duration

The results of the descriptive statistics and the effect size measures in Section 3.2.2. of this report show that there are differences in test durations between CBE and COLO. To further explore the differences using the mixed-effect modeling approach, Table 3.13 presents the details of the unconditional model, followed by more complex models with covariates and context effects. The corresponding R code is presented in Appendix A.

Table 3.13. Mixed-Effect Models used for Data Analysis—Test Duration

Model 1	Level 1: $Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 2	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{WinterRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 3	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{WinterRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + \beta_6 \mathrm{Altair_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 4	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{WinterRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + \beta_6 \mathrm{Altair_{ij}} + \beta_7 \mathrm{Altair_{ij}} * \mathrm{White_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 5	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{WinterRit_{ij}} + \beta_2 \mathrm{Time_{ij}} + \beta_3 \mathrm{Grade_{ij}} + \beta_4 \mathrm{White_{ij}} + \beta_5 \mathrm{Male_{ij}} + \beta_6 \mathrm{Altair_{ij}} + \beta_7 \mathrm{Altair_{ij}} * \mathrm{Male_{ij}} + \beta_8 \mathrm{Altair_{ij}} * \mathrm{White_{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + b_{0j} \mathrm{MalePct_j} + b_{1j} \mathrm{WhitePct_j} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$

Base models are as follows:

- *Model 1. One-way random effects ANOVA*. First, an unconditional model was fit in which school is a random effect. Students were nested within schools. Based on the results in Table 3.14, the correlation of the winter test duration among students within the same schools is approximately 0.14 in both content areas.

- *Model 2. Model with covariates except the Altair treatment effect.* Multiple covariates were added to account for variability in students' test durations, including students' winter RIT scores, time, grade, sex, race, and two school-level random coefficient effects: percentage of male (MalePct) and percentage of white (WhitePct) in schools. Similar to the adjustment for the variables in the RIT score models in Section 3.4.1. of this report, all variables other than the grade variable were grand-mean centered and the base-grade of the grade variable was readjusted to Grade 3.

Models with Altair effects are as follows:

- *Model 3. Model with the Altair treatment effect.* The comparison between Model 2 and Model 3 can help check the significance of Altair in predicting the winter test duration. In both content areas, the chi-square tests for deviance of Model 2 and Model 3 yield significant chi-squares (chi-squared = 42.57, p <0.0001 in Reading; chi-squared = 21.42, p <0.0001 in Mathematics).

- *Model 4. Model with Altair by race interaction effect.* The Altair by race interaction effect in Model 4 was further examined by adding the interaction variable to Model 3. The chi-square test for deviance of Model 3 and Model 4 yields a significant chi-square in Reading (chi-squared =9.55 , p < 0.01). The test was not significant in Mathematics (chi-squared = 0.13, p =0.72).

- *Model 5. Model with both Altair by race and Altair by sex interaction effects.* The chi-square test of Model 4 and Model 5 shows that the Altair and sex interaction is not significant in Reading (chi-squared =0.97, p =0.33) and the result is significant in Mathematics (chi-squared =4.44, p = 0.04).

Table 3.14. Random Effects—Test Duration

Model	Group	Name	Variance	SD	Correlation	
					(Intercept)	MalePct
Reading						
Model 1	School	(Intercept)	127.30	11.28	–	–
	Residual	–	790.10	28.11	–	–
Model 2	School	(Intercept)	98.86	9.94	–	–
		MalePct	146.02	12.08	-0.14	–
		WhitePct	31.47	5.61	0.49	0.65
	Residual	–	565.15	23.77	–	–
Model 3	School	(Intercept)	97.84	9.89	–	–
		MalePct	144.93	12.04	-0.12	–
		WhitePct	31.35	5.60	0.48	0.66
	Residual	–	565.02	23.77	–	–
Model 4	School	(Intercept)	97.98	9.90	–	–
		MalePct	145.17	12.05	-0.12	–
		WhitePct	30.98	5.57	0.48	0.63
	Residual	–	564.99	23.77	–	–

Model	Group	Name	Variance	SD	Correlation	
					(Intercept)	MalePct
Model 5	School	(Intercept)	97.98	9.90	–	–
		MalePct	145.17	12.05	-0.13	–
		WhitePct	30.99	5.57	0.48	0.63
	Residual	–	564.99	23.77	–	–
Mathematics						
Model 1	School	(Intercept)	114.90	10.72	–	–
	Residual	–	689.80	26.26	–	–
Model 2	School	(Intercept)	101.15	10.06	–	–
		MalePct	299.65	17.31	-0.25	–
		WhitePct	25.15	5.02	0.39	-0.02
	Residual	–	499.20	22.34	–	–
Model 3	School	(Intercept)	100.35	10.02	–	–
		MalePct	297.82	17.26	-0.25	–
		WhitePct	25.06	5.01	0.39	-0.02
	Residual	–	499.15	22.34	–	–
Model 4	School	(Intercept)	100.37	10.02	–	–
		MalePct	297.76	17.26	-0.25	–
		WhitePct	24.98	5.00	0.39	-0.03
	Residual	–	499.15	22.34	–	–
Model 5	School	(Intercept)	100.37	10.02	–	–
		MalePct	292.53	17.10	-0.26	–
		WhitePct	24.98	5.00	0.39	-0.04
	Residual	–	499.15	22.34	–	–

Table 3.15 presents the results of the fixed effect model estimates. As shown by the chi-square statistics in Table 3.16 that presents the model comparison results, Model 4 is the best-fitting model for Reading and Model 3 is the best-fitting model for Mathematics. However, as shown in Table 3.17, the model predictors account for 25% to 27% (SB estimator) of the variances in the data in both content areas. The added Altair effect in Model 3 makes trivial difference in the third decimal in the explained variance in both content areas. The f^2 value of any adjacent model is trivial. Thus, both the Altair variable and the Altair interaction variable do not contribute meaningfully to explain the variances in the test duration data, and they are negligible effects. Overall, it can be concluded that there is no Altair effect in the test duration data.

Table 3.15. Fixed Effects—Test Duration

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Reading										
(Intercept)	58.60	0.43	59.89	0.39	64.24	0.77	65.56	0.88	65.90	0.95
WinterRit	–	–	0.62	0.00	0.62	0.00	0.62	0.00	0.62	0.00
Time	–	–	-0.03	0.01	-0.03	0.01	-0.03	0.01	-0.03	0.01
Grade	–	–	1.93	0.06	1.90	0.06	1.90	0.06	1.90	0.06
White	–	–	-3.44	0.19	-3.44	0.19	-6.38	0.97	-6.36	0.97
Male	–	–	-3.67	0.12	-3.67	0.12	-3.66	0.12	-4.35	0.70
Altair	–	–	–	–	-4.41	0.68	-5.76	0.80	-6.11	0.88
Altair:White	–	–	–	–	–	–	3.02	0.98	3.00	0.98
Altair:Male	–	–	–	–	–	–	–	–	0.70	0.71
Mathematics										
(Intercept)	56.42	0.41	59.45	0.40	62.59	0.79	62.74	0.89	63.33	0.93
WinterRit	–	–	0.53	0.00	0.53	0.00	0.53	0.00	0.53	0.00
Time	–	–	-0.02	0.01	-0.02	0.01	-0.02	0.01	-0.02	0.01
Grade	–	–	1.88	0.06	1.87	0.06	1.87	0.06	1.87	0.06
White	–	–	-3.38	0.18	-3.38	0.18	-3.64	0.75	-3.63	0.75
Male	–	–	-6.62	0.12	-6.62	0.12	-6.62	0.12	-7.74	0.55
Altair	–	–	–	–	-3.27	0.71	-3.42	0.83	-4.04	0.88
Altair:White	–	–	–	–	–	–	0.28	0.77	0.26	0.77
Altair:Male	–	–	–	–	–	–	–	–	1.18	0.56

Table 3.16. Model Comparison Indexes—Test Duration

Models	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr(>Chisq)
Reading								
Model 2	13	1,524,874	1,525,004	-762,424	1,524,848	42.57	1.00	0.00
Model 3	14	1,524,833	1,524,974	-762,403	1,524,805			
Model 3	14	1,524,833	1,524,974	-762,403	1,524,805	9.55	1.00	0.00
Model 4	15	1,524,826	1,524,976	-762,398	1,524,796			
Model 4	15	1,524,826	1,524,976	-762,398	1,524,796	0.97	1.00	0.33
Model 5	16	1,524,827	1,524,987	-762,397	1,524,795			
Mathematics								
Model 2	13	1,478,660	1,478,790	-7.39E+05	1,478,634	21.42	1.00	0.00
Model 3	14	1,478,640	1,478,780	-7.39E+05	1,478,612			
Model 3	14	1,478,640	1,478,780	-7.39E+05	1,478,612	0.13	1.00	0.72
Model 4	15	1,478,642	1,478,792	-7.39E+05	1,478,612			
Model 4	15	1,478,642	1,478,792	-7.39E+05	1,478,612	4.44	1.00	0.04
Model 5	16	1,478,640	1,478,800	-7.39E+05	1,478,608			

Table 3.17. Results of the Variance Explained—Test Duration

Model	RB1	RB2	SB	MVP
Reading				
Model 2	0.285	0.222	0.276	0.363
Model 3	0.285	0.230	0.277	0.363
Model 4	0.285	0.229	0.277	0.363
Mathematics				
Model 2	0.276	0.118	0.254	0.365
Model 3	0.277	0.125	0.255	0.365
Model 4	0.277	0.125	0.255	0.365
Model 5	0.277	0.125	0.255	0.365

3.4.3. Test Length

The results of the descriptive statistics and the effect size measures in Section 3.2.2. of this report show that there are differences in test length between the two groups. This section further explores the differences using the mixed-effect modeling approach. Table 3.18 presents the list of models used for the analysis. The corresponding R code is presented in Appendix A.

Table 3.18. Mixed-Effect Models used for Data Analysis—Test Length

Model 1	Level 1: $Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 2	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{Grade}_{\mathrm{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$
Model 3	Level 1: $Y_{ij} = \beta_{0j} + \beta_1 \mathrm{Grade}_{\mathrm{ij}} + \beta_2 \mathrm{Altair}_{\mathrm{ij}} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$

The models are as follows:

- *Model 1. One-way random effects ANOVA*. An unconditional model is fit in which school is a random effect. Students were nested within schools.

- *Model 2. Model with covariates except Altair effect.* A grade covariate was added to account for variability in students' test lengths. The base grade of the grade variable was readjusted to Grade 3.

- *Model 3. Model with the Altair treatment effect.* The comparison between Model 2 and Model 3 can help determine the significance of Altair in predicting the winter test length. In both content areas, the chi-square tests for deviance of Model 2 and Model 3 yield significant chi-squares (as shown in Table 3.21).

Table 3.19 and Table 3.20 shows the estimates of the Random and the fixed effects of all three models for both content areas. In Reading, the estimated intercept of the third model is 41.25, which is a Grade 3 student's winter Reading test length. The coefficient of grade variable is -0.46, meaning that, on average, each additional grade increase is associated with a decrease of 0.46 test length on the winter test. The coefficient of Altair variable is 0.68, meaning that, on average, COLO students' test length is 0.68 items more than the CBE students'.

Table 3.19. Random Effects—Test Length

Model	Group	Name	Variance	SD
Reading				
Model 1	School	(Intercept)	0.30	0.55
	Residual	–	1.23	1.11
Model 2	School	(Intercept)	0.38	0.62
	Residual	–	0.78	0.88
Model 3	School	(Intercept)	0.38	0.62
	Residual	–	0.77	0.88
Mathematics				
Model 1	School	(Intercept)	3.27	1.81
	Residual	–	10.87	3.30
Model 2	School	(Intercept)	5.42	2.33
	Residual	–	5.12	2.26
Model 3	School	(Intercept)	5.45	2.33
	Residual	–	5.12	2.26

The model comparison results using the chi-square tests for deviance are shown in Table 3.21. The tests are statistically significant for both content areas, indicating that the models with the Altiar effect are better than the ones without the Altiar effect. However, as shown in Table 3.22, the model predictors account around 25% of the variances in the data in both content areas. The added Altair effect in Model 3 only makes a small difference in the third decimal in the explained variance in Reading, and no difference is detected in the third decimal in Mathematics from Model 2 to Model 3. The f^2 value of any adjacent models in both content areas is trivial. Thus, the Altair variable does not contribute meaningfully to explain the variances in students' winter test lengths, and they are negligible effects. Thus, it can be concluded that there is no Altair effect in the test length data.

Table 3.20. Fixed Effects—Test Length

	Model 1		Model 2		Model 3	
	Estimate	SE	Estimate	SE	Estimate	SE
Reading						
(Intercept)	40.67	0.02	41.27	0.02	41.25	0.02
Grade	–	–	-0.46	0.00	-0.46	0.00
Altair(COLO)	–	–	–	–	0.68	0.03
Mathematics						
(Intercept)	50.82	0.07	48.61	0.08	48.62	0.09
Grade	–	–	1.65	0.00	1.65	0.00
Altair(COLO)	–	–	–	–	-0.31	0.07

Table 3.21. Model Comparison Indexes—Test Length

Model	Df	AIC	BIC	logLik	Deviance	Chisq	Chi Df	Pr(>Chisq)
Reading								
Model 2	13	432,528	432,568	-216,260	432,520	711.18	1.00	< 2.2e-16
Model 3	14	431,819	431,869	-215,905	431,809			
Mathematics								
Model 2	13	733,634	733,674	-366,813	733,626	17.71	1.00	2.58E-05
Model 3	14	733,618	733,668	-366,804	733,608			

Table 3.22. Results of the Variance Explained—Test Length

Model	RB1	RB2	SB	MVP
Reading				
Model 2	0.367	-0.267	0.242	0.519
Model 3	0.369	-0.268	0.243	0.523
Mathematics				
Model 2	0.529	-0.662	0.254	0.609
Model 3	0.529	-0.669	0.252	0.608

3.5. Item Selection Process

3.5.1. Content Analysis

This section compares the test content of MAP Growth tests administered on CBE and COLO. Table 3.23 summarizes the total counts and percentages by grade and MAP Growth test. The counts are mostly evenly distributed across the given grades, with the exceptions being CBE Grade 2 in both content areas and COLO Grade 2 in Mathematics. Table 3.24 and Table 3.25 summarize the number of items on COLO and CBE. As shown in the tables, the number of items meets the constraints.

Table 3.23. Student Count by Grade and MAP Growth Test

Grade	Reading K–2		Reading 2–5		Reading 6+		Mathematics K–2		Mathematics 2–5		Mathematics 6+	
	N	%	N	%	N	%	N	%	N	%	N	%
COLO												
K	14,914	50.16	–	–	–	–	14,089	47.47	–	–	–	–
1	14,819	49.84	–	–	–	–	14,445	48.67	–	–	–	–
2	–	–	17,999	22.22	–	–	1,148	3.87	14,295	19.02	–	–
3	–	–	20,790	25.67	–	–	–	–	19,692	26.20	–	–
4	–	–	20,803	25.68	–	–	–	–	20,327	27.05	–	–
5	–	–	21,410	26.43	–	–	–	–	20,844	27.73	–	–
6	–	–	–	–	19,556	34.86	–	–	–	–	19,416	34.62
7	–	–	–	–	17,900	31.91	–	–	–	–	17,956	32.01
8	–	–	–	–	18,644	33.23	–	–	–	–	18,719	33.37
CBE												
K	91	57.59	–	–	–	–	595	52.52	–	–	–	–
1	67	42.41	–	–	–	–	526	46.43	–	–	–	–
2	–	–	62	2.78	–	–	12	1.06	712	17.65	–	–
3	–	–	392	17.57	–	–	–	–	1,083	26.85	–	–
4	–	–	977	43.79	–	–	–	–	1,222	30.29	–	–
5	–	–	800	35.86	–	–	–	–	1,017	25.21	–	–
6	–	–	–	–	1,095	40.24	–	–	–	–	1,142	40.41
7	–	–	–	–	786	28.89	–	–	–	–	819	28.98
8	–	–	–	–	840	30.87	–	–	–	–	865	30.61

Table 3.24. Item Calibration Status Counts

Calibration Status	Item Count	Reading K–2		Reading 2–5		Reading 6+		Mathematics K–2		Mathematics 2–5		Mathematics 6+	
		N	%	N	%	N	%	N	%	N	%	N	%
COLO													
FT	1	–	–	122	0.15	62	0.11	–	–	–	–	–	–
FT	2	–	–	–	–	1	0.00	–	–	–	–	–	–
FT	3	–	–	56	0.07	8	0.01	–	–	–	–	–	–
OP	37	–	–	39	0.05	3	0.01	–	–	–	–	–	–
OP	38	–	–	8	0.01	4	0.01	–	–	–	–	–	–
OP	39	–	–	120	0.15	57	0.10	–	–	–	–	–	–
OP	40	–	–	72,823	89.90	51,422	91.66	–	–	–	–	–	–
OP	41	–	–	2,581	3.19	1,530	2.73	–	–	–	–	–	–
OP	42	–	–	1,633	2.02	873	1.56	–	–	–	–	–	–
OP	43	29,733	100.00	3,798	4.69	2,211	3.94	29,682	100.00	–	–	–	–
OP	50	–	–	–	–	–	–	–	–	1,201	1.60	2,100	3.74
OP	51	–	–	–	–	–	–	–	–	7,097	9.44	10,184	18.16
OP	52	–	–	–	–	–	–	–	–	11,061	14.72	13,260	23.64
OP	53	–	–	–	–	–	–	–	–	55,799	74.24	30,547	54.46

Calibration Status	Item Count	Reading K–2		Reading 2–5		Reading 6+		Mathematics K–2		Mathematics 2–5		Mathematics 6+	
		N	%	N	%	N	%	N	%	N	%	N	%
CBE													
FT	1	–	–	622	27.88	–	–	–	–	–	–	–	–
OP	40	–	–	576	25.82	–	–	–	–	–	–	–	–
OP	41	–	–	1,533	68.71	2,544	93.50	–	–	–	–	–	–
OP	42	–	–	61	2.73	51	1.87	–	–	–	–	–	–
OP	43	158	100.00	61	2.73	126	4.63	1,133	100.00	–	–	–	–
OP	51	–	–	–	–	–	–	–	–	1,876	46.50	1,970	69.71
OP	52	–	–	–	–	–	–	–	–	474	11.75	253	8.95
OP	53	–	–	–	–	–	–	–	–	1,684	41.75	603	21.34

Table 3.25. Content Constraint and Guideline Results—Number of Items

MAP Growth Test	#Items*								
	Total			OP			FT		
	Constraint	COLO	CBE	Constraint	COLO	CBE	Constraint	COLO	CBE
Reading K–2	*43*	43	43	*40–43*	43	43	*0–3*	0	0
Reading 2–5	*40–43*	40–43	41–43	*36–43*	37–43	40–43	*0–4*	1–3	1
Reading 6+	*40–43*	40–43	41–43	*36–43*	37–43	41–43	*0–4*	1–3	0
Math K–2	*43*	43	43	*40–43*	43	43	*0–3*	0	0
Math 2–5	*50–53*	50–53	51–53	*47–53*	50–53	51–53	*0–3*	0	0
Math 6+	*50–53*	50–53	51–53	*47–53*	50–53	51–53	*0–3*	0	0

Table 3.26 and Table 3.27 summarize the instructional area item counts. All the CBE tests meet the constraints. The CBE tests' item count ranges are closed to the corresponding. Compared to the CBE tests, the COLO tests fell short on the minimum number of items per instructional goal. And the COLO's instructional goal level maximum item counts tend to be higher than the CBE's. CBE has better control of item balance by instructional areas because it allows configuration of constraints and guidelines in blueprint.

Table 3.26. Content Constraint and Guideline Results—Number of Items per Instructional Area

MAP Growth Test	#Items per Instructional Area					
	Min. #Items			Range #Items		
	Constraint	COLO	CBE	Guideline	COLO	CBE
Reading K–2	*10*	4	10	*10–13*	4–19	10–13
Reading 2–5	*7*	4	7	*7–15*	4–21	7–15
Reading 6+	*7*	4	7	*7–15*	4–19	7–15
Math K–2	*10*	6	10	*10–13*	6–18	10–13
Math 2–5	*11*	7	11	*12–17*	8–20	11–20
Math 6+	*11*	8	11	*12–17*	11–20	11–20

Table 3.27. Instructional Area Item Counts

Instructional Area	Engine	Min.	Max.
Reading K–2			
Comprehension	COLO	8	16
Concepts of Print, Phonological Awareness, Word Analysis	COLO	4	19
Vocabulary	COLO	7	16
Writing	COLO	4	19
Comprehension	CBE	10	13
Concepts of Print, Phonological Awareness, Word Analysis	CBE	10	13
Vocabulary	CBE	10	13
Writing	CBE	10	13
Reading 2–5			
Build and Use Vocabulary	COLO	5	21
Informational Text: Characteristics of Text	COLO	5	13
Informational Text: Main Idea and Analysis	COLO	4	13
Literary Text: Characteristics of Text	COLO	5	15
Literary Text: Theme and Analysis	COLO	5	14
Build and Use Vocabulary	CBE	7	15
Informational Text: Characteristics of Text	CBE	7	14
Informational Text: Main Idea and Analysis	CBE	7	14
Literary Text: Characteristics of Text	CBE	7	14
Literary Text: Theme and Analysis	CBE	7	15
Reading 6+			
Build and Use Vocabulary	COLO	4	19
Informational Text: Characteristics of Text	COLO	5	13
Informational Text: Main Idea and Analysis	COLO	5	13
Literary Text: Characteristics of Text	COLO	5	18
Literary Text: Theme and Analysis	COLO	5	14
Build and Use Vocabulary	CBE	7	15
Informational Text: Characteristics of Text	CBE	7	15
Informational Text: Main Idea and Analysis	CBE	7	15
Literary Text: Characteristics of Text	CBE	7	15
Literary Text: Theme and Analysis	CBE	7	15
Mathematics K–2			
Algebra	COLO	6	15
Data	COLO	8	15
Geometry	COLO	7	16
Number	COLO	8	18
Algebra	CBE	10	13
Data	CBE	10	13
Geometry	CBE	10	13
Number	CBE	10	13

Instructional Area		Engine	Min.	Max.
Mathematics 2–5				
	Algebra	COLO	7	17
	Data		12	18
	Geometry		11	18
	Number		12	20
	Algebra	CBE	11	19
	Data		11	20
	Geometry		11	19
	Number		11	19
Mathematics 6+				
	Algebra	COLO	9	20
	Data		12	18
	Geometry		11	18
	Number		8	20
	Algebra	CBE	11	19
	Data		11	20
	Geometry		11	19
	Number		11	19

Table 3.28, Table 3.29, and Table 3.30 summarize the passage counts and passage lengths in the tests. All the CBE tests meet the constraints. Most students received two passages with four items on COLO. On CBE, most students received either zero or one passages, and only between 15–30% of the students received either two or three passages. Most CBE passages have three items. The results also show that CBE and COLO are comparable in terms of item positions, with CBE passage item positions ranging from 7–21 and COLO passage item positions ranging from 7–18.

Table 3.28. Content Constraint and Guideline Results—Passages

MAP Growth Test	#Passages			#Items per Passage		
	Constraint	COLO	CBE	Constraint	COLO	CBE
Reading 2–5	*0–3*	0–3	0–3	*3–5*	1–5	3–5
Reading 6+	*0–3*	0–3	0–3	*3–5*	1–5	3-5

Table 3.29. Percent of Students Receiving a Specific Number of Passages

MAP Growth Test	Engine	%Students Receiving Each #Passages				Passage Item Position
		0	1	2	3	
Reading 2–5	COLO	18.79	21.00	60.15	0.06	[7, 8,9,10,11,12,13,14,15,16,17,18]
	CBE	36.17	34.2	17.12	12.51	[7, 8,9,10,11,12,13,14,15,16,17,18, 19, 20]
Reading 6+	COLO	8.91	22.08	69.00	0.01	[7, 8,9,10,11,12,13,14,15,16,17,18]
	CBE	39.58	44.65	12.42	3.34	[7, 8,9,10,11,12,13,14,15,16,17,18, 19, 20, 21]

Table 3.30. Percent of Students Receiving a Specific Number of Items in a Passage

MAP Growth Test	Engine	%Students Receiving #Items in a Passage				
		1	2	3	4	5
Reading 2–5	COLO	0.97	0.77	3.14	95.02	0.11
	CBE	–	–	95.47	4.10	0.42
Reading 6+	COLO	2.05	1.65	18.22	78.00	0.07
	CBE	–	–	96.16	3.42	0.42

3.5.2. Item Exposure

Table 3.31 summarizes the item exposure rates for each test during the Winter 2020 administration. For example, an item exposure rate of 0.1 indicates an item was exposed to 10% of the test cases, whereas an item exposure rate of 0 indicates an item was not selected in any test case. In general, more Reading items were not exposed by CBE than by COLO, and the exposure rates are comparable in Mathematics tests in both groups. Most of the exposed items have an exposure rate between 0–10%, and very few items were exposed more than 10%. None of the Reading and Mathematics items were exposed more than 40%.

A closer look at the Reading items that were not exposed to the students on CBE shows that all the 15% and 20% unexposed items in Reading 2–5 and Reading 6+, respectively, are passage-related items, which is probably because CBE has a very different passage selection algorithm than COLO. CBE selects the most informative items given the constraints on the number of unique passage IDs permitted and the minimum and/or maximum number of items that use a given passage, whereas COLO selects the passage based on mean difficulty of all items within it and then the most informative items in that passage.

Table 3.31. Item Exposure Rates

MAP Growth Test	Exposure Rate	#Items by Mode			
		COLO*		CBE	
		N	%	N	%
Reading K–2	0	1	–	962	33
	(0.0, 0.1]	2,935	100	1,946	67
	(0.1, 0.2]	10	–	15	1
	(0.2, 0.3]	1	–	–	–
	(0.3, 0.4]	–	–	–	–
	(0.4, 0.5]	–	–	–	–
	>0.5	–	–	–	–
Reading 2–5	0	10	–	532	15
	(0.0, 0.1]	3,478	100	2,940	85
	(0.1, 0.2]	–	–	2	–
	(0.2, 0.3]	–	–	1	–
	(0.3, 0.4]	–	–	–	–
	(0.4, 0.5]	–	–	–	–
	>0.5	–	–	–	–

MAP Growth Test	Exposure Rate	#Items by Mode			
		COLO*		CBE	
		N	%	N	%
Reading 6+	0	44	1	726	20
	(0.0, 0.1]	3,689	99	2,989	80
	(0.1, 0.2]	3	–	–	–
	(0.2, 0.3]	–	–	–	–
	(0.3, 0.4]	–	–	–	–
	(0.4, 0.5]	–	–	–	–
	>0.5	–	–	–	–
Mathematics K–2	0	–	–	45	3
	(0.0, 0.1]	1,725	98	1,652	94
	(0.1, 0.2]	18	1	47	3
	(0.2, 0.3]	8	–	14	1
	(0.3, 0.4]	7	–	–	–
	(0.4, 0.5]	–	–	–	–
	>0.5	–	–	–	–
Mathematics 2–5	0	–	–	–	–
	(0.0, 0.1]	2,819	98	2,826	98
	(0.1, 0.2]	64	2	57	2
	(0.2, 0.3]	–	–	–	–
	(0.3, 0.4]	–	–	–	–
	(0.4, 0.5]	–	–	–	–
	>0.5	–	–	–	–
Mathematics 6+	0	–	–	59	1
	(0.0, 0.1]	5,114	100	5,055	99
	(0.1, 0.2]	–	–	–	–
	(0.2, 0.3]	–	–	–	–
	(0.3, 0.4]	–	–	–	–
	(0.4, 0.5]	–	–	–	–
	> 0.5	–	–	–	–

*The COLO item pools might have been changed when the tests were administered to the students. However, the changes shall be small, and we assume that the majority of the item pool stay the same.

While there are no passage-related items in Reading K–2, the CBE students' momentary ability distribution was compared with the Reading K–2 item pool item difficulty distribution. Table 3.32 presents the results of the comparison. "Momentary Theta" is the students' momentary RIT scores transformed back to the theta scale. For example, the results in the third row show that there are two momentary thetas of all the CBE Reading K–2 test cases and five Reading K–2 item difficulties in the range of (-9,-8.5], and 60% of the items in the range were not exposed. There are more items than students at the lower scale (between -10 to -6), which caused the low item pool use rate. It can be further explained by the fact that many less students participated in the CBE tests than COLO (as shown in Table 3.2). The difference between the theta and the item difficulty distributions and the small CBE sample size indicate that there are not have enough students compared to the number of items available at each bin at the lower scale, which caused the low use rate of the Reading K–2 item pool.

Table 3.32. Reading K–2 Momentary Theta and Item Difficulty Distributions

Scale	#Momentary Theta	#Item Difficulty	%Items not Exposed
(-10,-9.5]	–	–	–
(-9.5,-9]	1	–	–
(-9,-8.5]	2	5	60.00
(-8.5,-8]	6	14	71.43
(-8,-7.5]	20	44	63.64
(-7.5,-7]	65	87	56.32
(-7,-6.5]	74	150	62.00
(-6.5,-6]	166	186	63.98
(-6,-5.5]	371	232	53.02
(-5.5,-5]	568	273	50.92
(-5,-4.5]	583	281	38.79
(-4.5,-4]	649	295	29.15
(-4,-3.5]	838	253	15.42
(-3.5,-3]	790	284	17.61
(-3,-2.5]	643	219	10.05
(-2.5,-2]	597	186	12.90
(-2,-1.5]	431	128	16.41
(-1.5,-1]	231	98	28.57
(-1,-0.5]	179	77	12.99
(-0.5,0]	187	47	14.89
(0,0.5]	109	36	5.56
(0.5,1]	41	20	–
(1,1.5]	26	8	–
(1.5,2]	5	–	–
(2,2.5]	2	–	–
(2.5,3]	–	–	–
(3,3.5]	2	–	–
(3.5,4]	–	–	–

3.5.3. Engine Adaptivity

The engine's adaptivity was assessed via the delta value that indicates the difference between the item difficulty and students' momentary RIT. Figure 3.7 and Figure 3.8 illustrate the absolute delta for COLO and CBE in Reading and Mathematics, respectively. The y-axis is the mean absolute delta, and the x-axis represents the theta points at the percentile level. For example, the first x-axis point "1" in the top panel of Figure 3.7 represents the 10th percentile of the momentary RITs of all the Reading K–2 test cases. Overall, the absolute delta values for CBE are smaller than those for COLO, and all absolute delta values are less than 6 (on the RIT scale). This suggests the items selected by CBE are more adaptive than those by COLO, especially for students with very low or high achievement. However, COLO and CBE both show less adaptivity at the extremes.

Figure 3.7. Absolute Delta by RIT Percentile—Reading

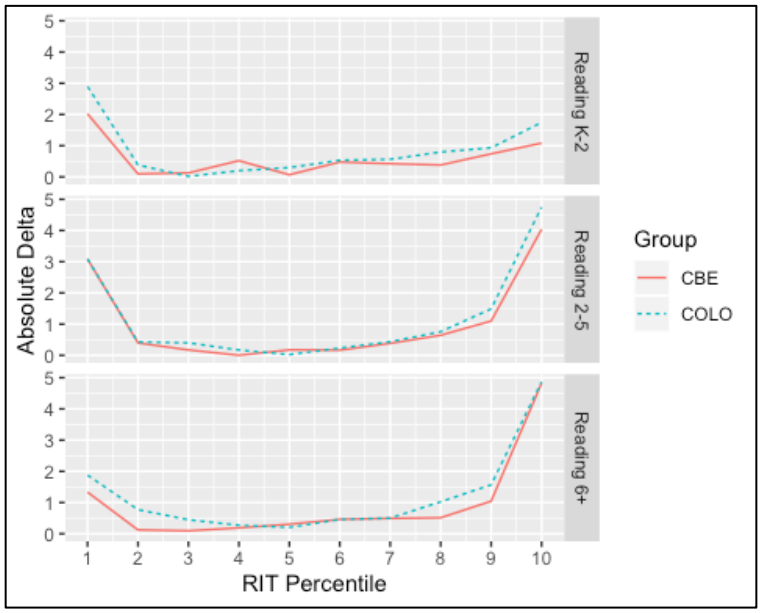
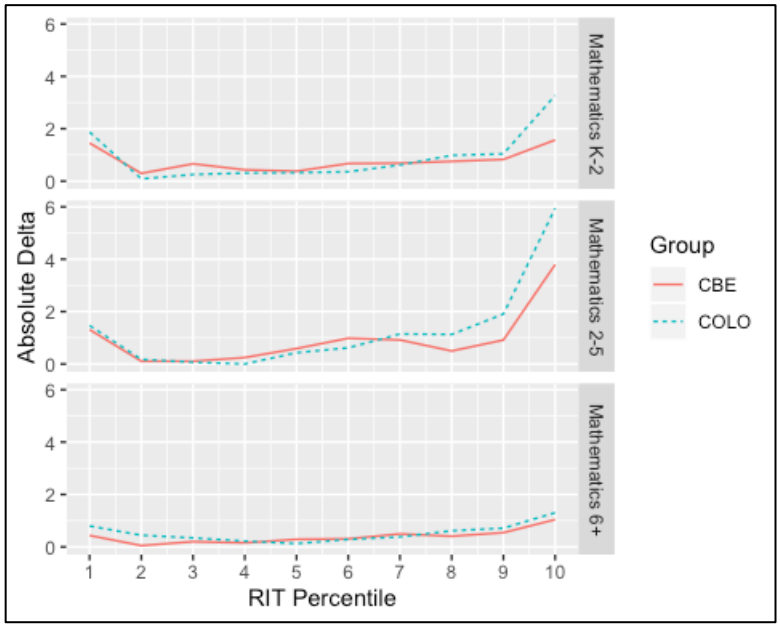


Figure 3.8. Absolute Delta by RIT Percentile—Mathematics



4. Conclusion

This study examined the comparability of the MAP Growth Reading and Mathematics tests administered on COLO and CBE from students in Nebraska in Fall 2019 and Winter 2020. Three major criteria (i.e., validity, psychometrics and reliability, and statistical assumption/test administration) were posed regarding the comparability of the tests administered on the two administration modes. Overall, the results of this study indicate that MAP Growth tests administered on COLO and CBE are comparable. Despite the few statistically significant differences across the two administration modes in RIT scores, test length, and test duration, a deeper dive of the analysis using a multilevel statistical model approach shows that the differences are not practically significant and the "Altair" treatment effect makes trivial differences in the variance explained in the variable of interest (e.g., students' winter RIT scores).

While the tests administered on CBE have higher precisions than those administered on COLO, the difference is small with a maximum value of 0.1. CBE also shows better adaptivity than COLO, especially for extremely low or high achievement students. The capability of CBE in providing higher score precision and better adaptivity are also confirmed by the Project Altair simulation study (Hu et al., 2020). Results suggest that the CBE tests could be one to two items shorter than the COLO tests without loss of precision.

The content analysis shows that the tests delivered on CBE meet all the constraints, including the total test lengths and number of operational and field test items, the passage-related constraints, and the instructional-level item count constraints. Most items on both the CBE and COLO tests have a less than 10% item exposure rate, and none of the items on both platforms have item exposure rates higher than 40%. The COLO Reading item use rate is higher than the CBE rate, but the item use rates are comparable across the two platforms for Mathematics. The difference between the item use rate in Reading across the two engines was due to (1) the differences in the passage item selection algorithm and (2) the CBE sample size being much smaller than COLO's. This indicates that there are more items than students at the lower scale. The CBE passage item selection algorithm focuses on selecting the most informative passage items first, whereas COLO focuses on selecting the passage based on the mean difficulty of all the items. The differences resulted in a much lower item use rate of the passage items in the CBE tests than the COLO tests, indicating that most of the passage items are not very informative. Thus, more informative passage items should be developed to deepen the MAP Growth item pool.

Overall, the empirical data mode comparability study reaches the same conclusions as the simulation study (Hu et al., 2020) that the two engines are comparable. However, while both engines deliver a reliable and valid test, CBE can maximize the flexibility in meeting various content requirements, has better adaptivity in picking items that are closer to students' momentary abilities, and provides slightly more precise scores.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergstrom, B. (1992, April*). Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Eignor, D. R., & Schaeffer, G. A. (1995, April). *Comparability studies for the GRE General CAT and the NCLEX using CAI*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Grund, S., Robitzsch, A., & Lüdke, O. (2016). *Package 'mitml'*. https://cran.r-project.org/web/packages/mitml/.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications (2nd ed.)*. New York, NY: Routledge.

Hu, A., Chien, M., & Meyer, P. (2020). *Comparability of MAP Growth tests administered through different technology and psychometric infrastructure: A simulation study*. Portland, OR: NWEA.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics,* 44, 1086–1105. http://dx.doi.org/10.1080/02664763.2016.1193725

Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781849209366

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods,* 17, 433– 451. http://dx.doi.org/10.1177/109442 8114541701

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and P&P edwational and psychological tests. A review of the literature* (College Board Report 88-8). New York College Entrance Examination Board.

McCoach, D., & Black, A. (2008). Assessing model adequacy. In A. O'Connell & D. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte, NC: Information Age.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and P&P cognitive ability tests: A meta-analysis. *Psychological Bulletin,* 114, 449–458.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution,* 4, 133–142. http://dx.doi.org/10.1111/j.2041- 210x.2012.00261.x

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K–3 reading tests. *Journal of Educational Computing Research, 32*(2), 153–166. https://doi.org/10.2190%2FD2HU-PVAW-BR9Y-J1CL

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models (2nd ed.)*. Thousand Oaks, CA.: Sage.

Recchia, A. (2010). R-squared measures for two-level hierarchical linear models using SAS. *Journal of Statistical Software, 32*, 1–9.

Roberts, J. K., Monaco, J. P., Stovall, H., & Foster, V. (2011). Explained variance in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 219–230). New York, NY: Routledge.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 2*, 233–247.

Segall, D. O. (1995, April). *Equating the CAT-ASV'AB: Experiences and lessons learned.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.)*. London, UK: Sage.

Tsai, T.-H., & Shin, C. D. (2013). A score comparability study for the NBDHE: Paper–pencil versus computer versions. *Evaluation & the Health Professions, 36*(2), 228–239. https://doi.org/10.1177/0163278712445203

Wang, T., & Kolen, M. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49. www.jstor.org/stable/1435437

Wang, J., Xie, H., & Fisher, J. H. (2011). *Multilevel models: Applications using SAS.* Göttingen, Germany: Walter de Gruyter. http://dx.doi.org/10.1515/9783110267709

Wright, B. D. (1999). Rasch measurement models. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 85–97). New York, NY: Pergamon.

Appendix A: R-Code of Mixed-Effect Models

Table A.1. lmer Model Syntax—Predicting Winter RIT Scores

NO.	Model	lmer Model Syntax
1	$Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_rit \sim 1 + (1 \mid SCHOOL)$
2	$Y_{ij} = \beta_{0j} + \beta_1 FallRit_{ij} + \beta_2 Time_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_rit \sim fall\_ritCentered + TimeDiffDaysCentered + GradeCentered + white + male + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
3	$Y_{ij} = \beta_{0j} + \beta_1 FallRit_{ij} + \beta_2 Time_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_rit \sim fall\_ritCentered + TimeDiffDaysCentered + GradeCentered + white + male + altair + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
4	$Y_{ij} = \beta_{0j} + \beta_1 FallRit_{ij} + \beta_2 Time_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + \beta_7 Altair_{ij} * Male_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_rit \sim fall\_ritCentered + TimeDiffDaysCentered + GradeCentered + white + altair * male + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
5	$Y_{ij} = \beta_{0j} + \beta_1 FallRit_{ij} + \beta_2 Time_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + \beta_7 Altair_{ij} * Male_{ij} + \beta_8 Altair_{ij} * White_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_rit \sim fall\_ritCentered + TimeDiffDaysCentered + GradeCentered + altair * white + altair * male + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$

Table A.2. lmer Model Syntax—Predicting Winter Test Duration

NO.	Model	lmer Model Syntax
1	$Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testDuration \sim 1 + (1 \mid SCHOOL)$
2	$Y_{ij} = \beta_{0j} + \beta_1 WinterRit_{ij} + \beta_2 TimeDiffDays_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testDuration \sim winter\_ritCentered + TimeDiffDaysCentered + GradeCentered + white + male + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
3	$Y_{ij} = \beta_{0j} + \beta_1 WinterRit_{ij} + \beta_2 TimeDiffDays_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testDuration \sim winter\_ritCentered + TimeDiffDaysCentered + GradeCentered + white + male + altair + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
4	$Y_{ij} = \beta_{0j} + \beta_1 WinterRit_{ij} + \beta_2 TimeDiffDays_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + \beta_7 Altair_{ij} * White_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testDuration \sim winter\_ritCentered + TimeDiffDaysCentered + GradeCentered + male + altair * white + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$
5	$Y_{ij} = \beta_{0j} + \beta_1 WinterRit_{ij} + \beta_2 TimeDiffDays_{ij} + \beta_3 Grade_{ij} + \beta_4 White_{ij} + \beta_5 Male_{ij} + \beta_6 Altair_{ij} + \beta_7 Altair_{ij} * Male_{ij} + \beta_8 Altair_{ij} * White_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + b_{0j} MalePercentage_j + b_{1j} WhitePercentage_j + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testDuration \sim winter\_ritCentered + TimeDiffDaysCentered + GradeCentered + altair * white + altair * male + (male.percentageCentered + white.percentageCentered \mid SCHOOL)$

Table A.3. lmer Model Syntax—Predicting Winter Test Length

NO.	Model	lmer Model Syntax
1	$Y_{ij} = \beta_{0j} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testLength \sim 1 + (1 \mid SCHOOL)$
2	$Y_{ij} = \beta_{0j} + \beta_1 Grade_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testLength \sim GradeCentered + (1 \mid SCHOOL)$
3	$Y_{ij} = \beta_{0j} + \beta_1 Grade_{ij} + \beta_2 Altair_{ij} + e_{ij}$, $e_{ij} \sim N(0, \sigma^2)$, $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\mu_{0j} \sim N(0, \tau_{00}^2)$	$win\_testLength \sim GradeCentered + altair + (1 \mid SCHOOL)$