March 2021

Office of Shared Accountability

MONTGOMERY COUNTY PUBLIC SCHOOLS, ROCKVILLE, MARYLAND

Student Outcomes on MAP Growth: Comparison of Virtual and In-Person Administrations

PREPARED BY:

Syretta R. James, Ph.D.
Shihching Jessica Liu, M.A.
Nyambura Maina, Ph.D.
Julie Wade, M.A.
Helen Wang, Ph.D.
Heather Wilson, Ph.D.
Natalie Wolanin, M.Ed.

ROCKVILLE, MARYLAND

850 Hungerford Drive
Rockville, Maryland 20850
301-740-3000

Dr. Jack R. Smith
Superintendent of Schools

Dr. Janet S. Wilson
Chief, Teaching, Learning & Schools

Dr. Kecia L. Addison
Director, Office of Shared Accountability

Published for the Office of Shared Accountability

Table of Contents

Executive Summary 1
Introduction 3
Test Duration and Mean MAP-M/R Scores in Virtual vs. In-person Setting 4
Reliability of MAP-Mathematics and MAP-Reading Test Scores: Fall 2016 to Fall 2020 15
Differences in the Conditional Growth Index (CGI) Between In-Person and Virtual Test Settings 20
References 30

Executive Summary

The impact of the COVID-19 pandemic continues to overwhelm the functioning and outcomes of educational systems throughout the nation. Since the beginning of the crisis, school systems have attempted to establish norms for monitoring student progress with assessments administered virtually. However, many stakeholder groups continue to express concern about the reliability of progress monitoring measures implemented in a virtual setting. The current analytic report addresses these concerns by examining the comparability of in-person and virtual test performance across a sample of Montgomery County Public Schools (MCPS) students who took the MAP Growth (MAP) assessment in both the in-person and virtual settings. Data from this analysis supports the continued use of MAP assessment data as a reliable measure of student performance regardless of the assessment setting. In addition, relatively consistent psychometric characteristics and trends were reported. These findings are consistent with recent national studies that evaluated the comparability of MAP Growth scores across different modes of implementation (Kuhfeld et al., 2020b; Meyer, 2020).

Key Findings

There is little difference in MCPS student MAP Growth performance among students tested in a virtual setting compared to an in-person setting.

- Trend analysis revealed consistency in MAP Growth scores, as measured by MAP Reading or Mathematics (MAP-R or MAP-M) Rasch Unit Scale (RIT) and the Conditional Growth Index (CGI), for students tested in both the in-person and virtual settings.
- In addition, the MAP-R/M RIT scores for all MCPS students remained consistent across testing settings.
- Although grade level differences exist in terms of RIT score distribution (in some cases scores decreased and in others the scores increased), the size of the differences were small indicating the magnitude of the effects on student performance was negligible.
- The average change in MAP-R growth, as measured by CGI, was lower in the virtual test setting when compared to the in-person setting. However, the change was small and generally fell within the range of expected normal growth. It is important to note, MAP scores are sensitive enough to respond to the subtle changes that might occur when there is a change in test setting.

Although there are observed differences in test duration for students who took MAP-R/M assessments in different settings, the differences are small and do not provide an indication that one setting is better than the other.

- For MAP-R testing, students took less time in the virtual vs. the in-person setting. However, there were little differences in MAP-R scores changes.

- On MAP M testing, differences in test duration were largely dependent on grade. Students in Grades 2, 4, and 5 evidenced the smallest negligible differences in test duration.
- In general, students receiving special education services and those identified as Limited English Proficient (LEP) did not spend a longer amount of time testing in the virtual setting. However, there were grade-level differences that suggest the qualitative differences in test duration observed among certain students groups is largely attributable to individual student-driven or technology-driven factors (e.g., motivation toward testing, technical difficulties, etc.) as opposed to the test setting.

Regardless of the test setting, data obtained from the MAP Growth assessment was found to be a consistent measure of student achievement in the areas of reading and mathematics.

- On measures of reliability and internal consistency (Standard Error of Measurement SEM) on the MAP-R/M assessments all findings indicate these measures provide an accurate picture of student performance regardless of the test setting.
- The satisfactory reliability observed for the fall 2020 test administration is corroborated by a study reporting high marginal reliability and test-retest reliability for virtual and in-person administrations of MAP in districts across the country (Kuhfeld et al., 2020). Taken together, these measures of reliability, consistent with measures of previous years, suggest that MCPS educators can have confidence in results obtained from the fall 2020 MAP administration.

Introduction

The impact of the COVID-19 pandemic continues to overwhelm the functioning and outcomes of educational systems throughout the nation. The public education system is under particular scrutiny given that students, families, and educators are under considerable stress to maintain academic progress. Since the beginning of the crisis, school-systems have attempted to establish norms for monitoring student progress with assessments administered virtually. However, many stakeholder groups expressed concern about the reliability of assessments implemented in a virtual setting. While a few studies have provided strong support for the continued use of nationally normed performance measures such as the MAP assessment for progress monitoring (Kuhfeld et al., 2020b; Meyer, 2020), local educators continue to be reluctant to support test data that comes out of their individual school districts. Therefore, more information is needed to quell educator suspicions and shape their perspectives using data obtained from their local school districts.

The current report aims to address educator concerns by providing a direct look at Montgomery County Public Schools' (MCPS) student performance data related to MAP assessments. The report is arranged under three sections including:

1) Test Duration and Mean MAP-M/R Scores in Virtual vs. In-person Setting
2) Reliability of MAP-Mathematics and MAP-Reading Test Scores: Fall 2016 to Fall 2020
3) Differences in the Conditional Growth Index (CGI) Between In-Person and Virtual Test Settings

The data obtained in this report should be used to aid in the discussion about the utility of using MAP-R/M assessments as reliable tools in assessing student progress during and after the pandemic. For schools seeking additional guidance and support on ways to create similar testing environments in both virtual and in-person settings, the following resource is offered: https://nwea.force.com/nweaconnection/s/remote-testing-resources?language=en_US.

Test Duration and Mean MAP-M/R Scores in Virtual vs. In-person Setting

Helen Wang, Ph.D.
Natalie Wolanin, M.Ed.

Purpose and Data Sources

The purpose of this analysis was to examine differences in test duration and mean MAP RIT scores associated with the assessment setting (in-person vs. virtual). Data obtained from the fall 2019 MAP administration was used as the measurement of in-person test performance. These data were compared to data obtained from the fall 2020 MAP administration, which was used to measure virtual test performance. For data obtained during the fall 2020 MAP administration, it is worthy to note that the assessment setting reflected not only the method by which the assessment was administered (in-person vs. virtual), but also the way students received instruction (in-person vs. virtual) since students received virtual only instruction beginning in March 2020. For the purpose of this analysis, data were disaggregated by grade level and service groups (special education and LEP).

Test Duration Analysis

One aspect of examining differences between the in-person and virtual setting is comparing test durations. Figures 1 through 6 display the average minutes to complete the MAP assessments by setting (in-person vs. virtual) and by grade level for all students, students receiving special education services and those identified as LEP. It is important to note that students receiving special education services typically receive accommodations (i.e., extended time, access to readers, calculators, etc.) during testing. However, the current analysis did not assess the degree to which accommodations might have influenced outcomes regardless of the administration method.

Test Duration for MAP-M

In general, students who took MAP-M virtually took a longer time when compared to the in-person administration (Figure 1). However, the differences in test duration were noticeably small and varied among Grades levels. Importantly, the smallest difference was observed in Grades 2, 4, and 5. In addition, among Grade 4 and 5 students receiving special education services and those identified as LEP who took MAP-M, the average minutes to test was slightly less virtually than in with in-person administration (Figures 2 and 3). It is important to note that there were known technical difficulties with access to the fall 2020 Northwest Evaluation Association (NWEA) assessment administered virtually. It is unclear how much these technical issues influenced the time to take MAP-M assessments in the virtual setting.

Figure 1

Mean MAP-M Test Duration for All Students by Grade by Setting (In-Person vs. Virtual)

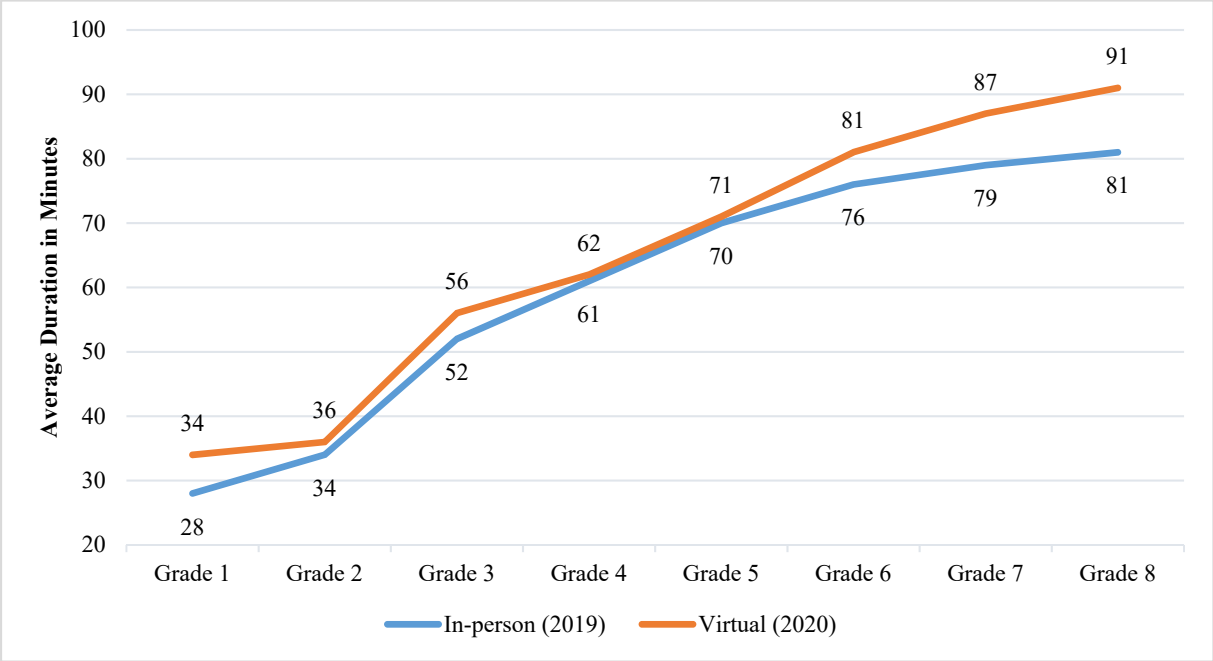


Figure 2

Mean MAP-M Test Duration for Student Receiving Special Education Services by Grade by Setting (In-Person vs. Virtual)

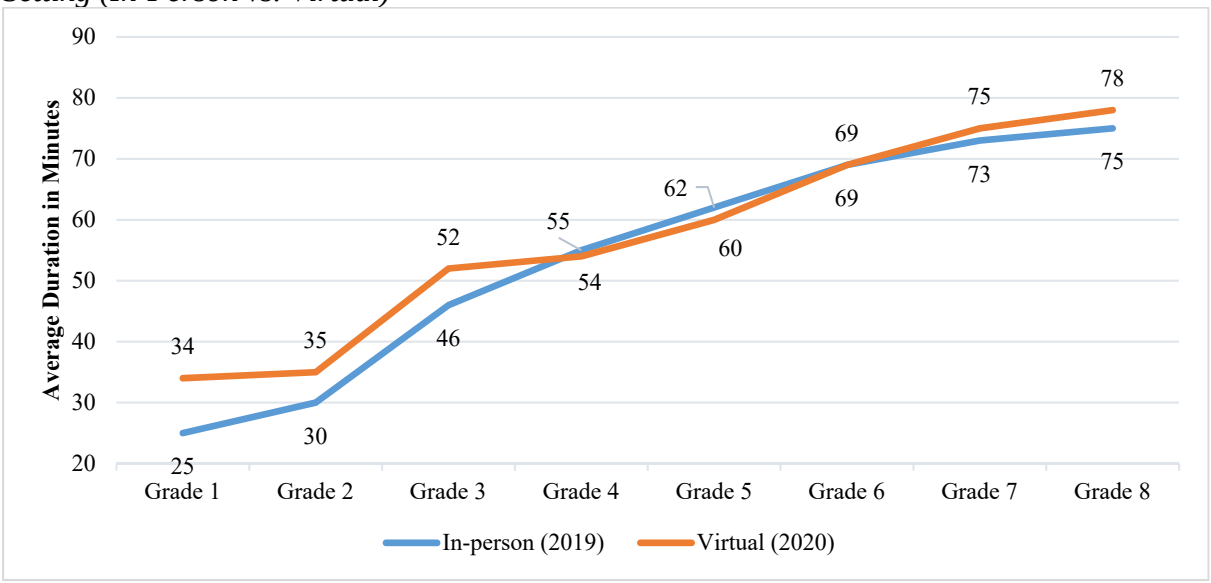
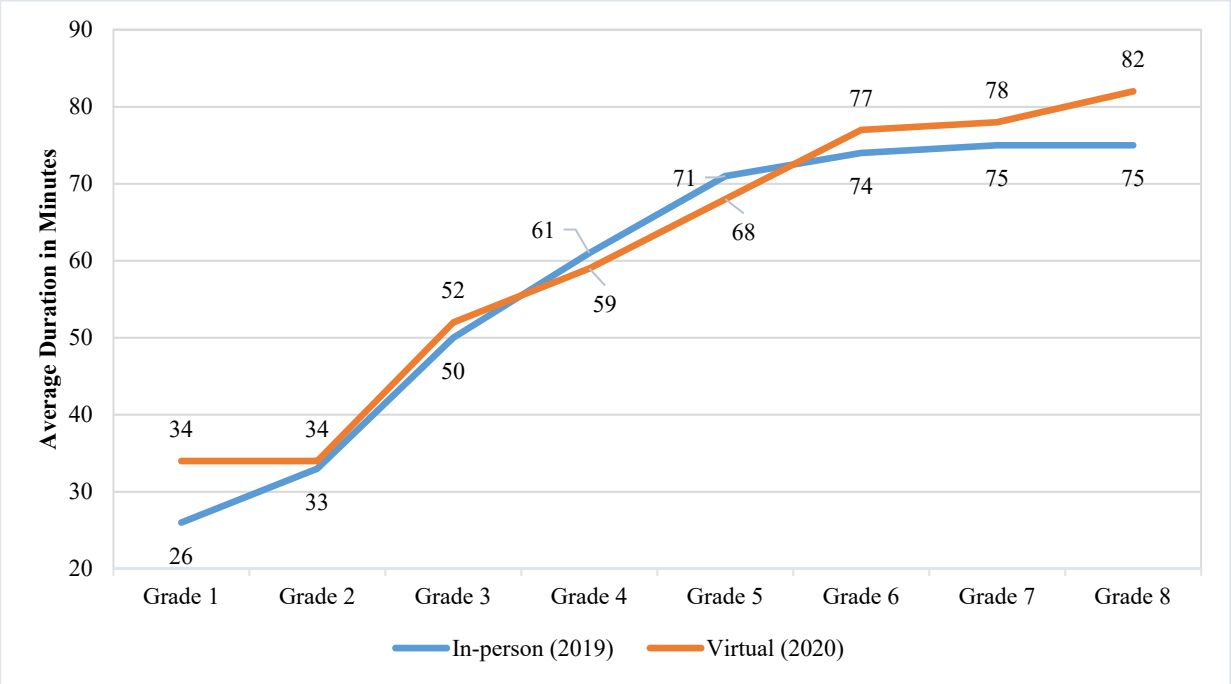


Figure 3

Mean MAP-M Test Duration for Students Identified as LEP by Grade by Setting (In-Person vs. Virtual)



Test Duration for MAP-R

Unlike during MAP-M testing, students who took MAP-R used a shorter amount of time for the virtual administration when compared to in-person testing. However, the differences in test duration were small yielding an average of 5 minutes of difference between the two settings. The smallest difference in testing time was observed in students in Grades 7 and 8 (Figure 4). In general, students receiving special education services evidenced a shorter MAP-R test duration in the virtual setting, with the exception being students in Grade 3 (Figures 5 and 6). The biggest differences in test duration were observed among students identified as LEP. Again, it is unclear if any of the assessment outcomes related to test duration were impacted by the noted technical issues virtual test takers experienced during the fall 2020 MAP-R administration.

Figure 4

Mean MAP-R Test Duration for All Students by Grade by Setting (In Person vs. Virtual)

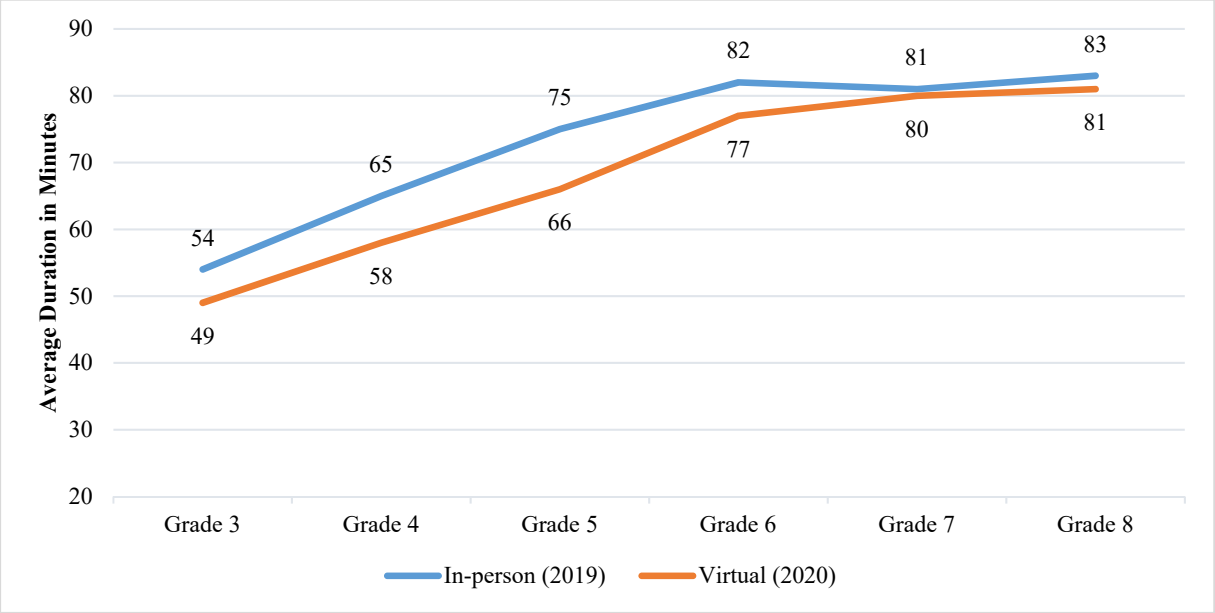


Figure 5

Mean MAP-R Test Duration for Students Receiving Special Education Services by Grade by Setting (In Person vs. Virtual)

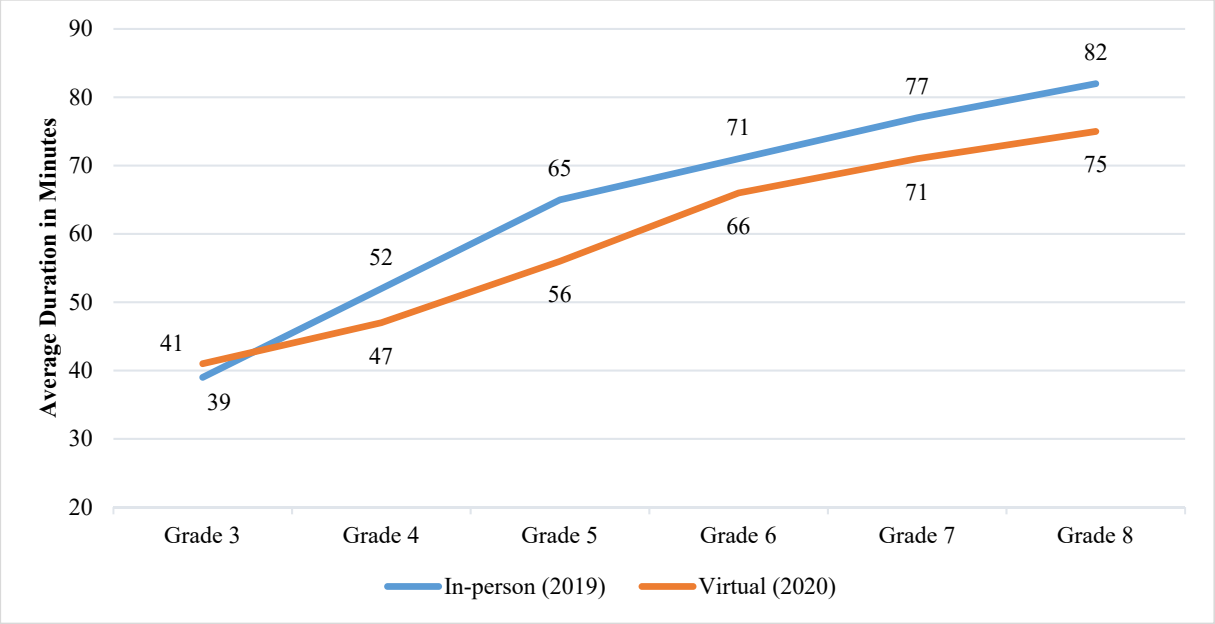
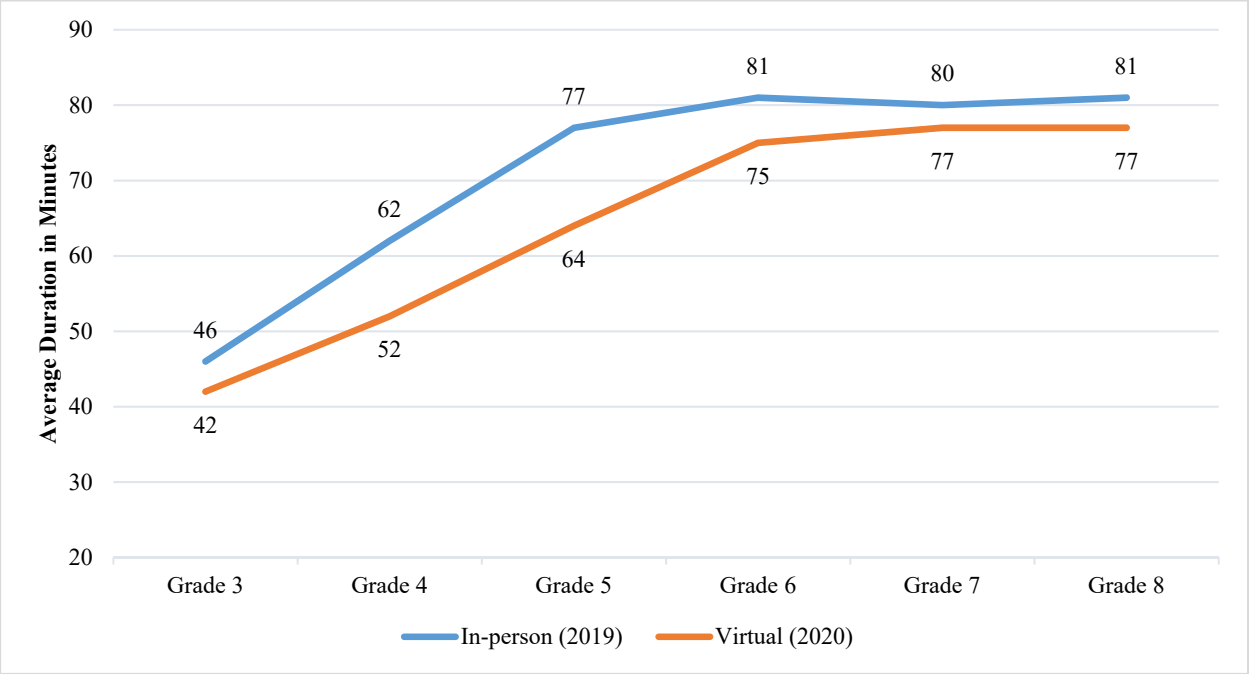


Figure 6

Mean MAP-R Test Duration by Students Identified as LEP by Grade by Setting (In Person vs. Virtual)



Analysis of Performance Differences in Virtual and In-Person MAP RIT Scores

To further examine how test setting impacts MAP performance data, an analysis of covariance (ANCOVA) statistical procedure was conducted to examine mean MAP-R/M RIT score differences in Grades 1 through 8 for MAP-M and Grades 3 through 8 for MAP-R. Using the ANCOVA procedure allows the examination of mean differences, while controlling for the effects associated with extraneous factors such as student characteristics (race/ethnicity, Free and Reduced-price Meals System (FARMS), LEP, and Special Education) and prior achievement (prior year's fall RIT score). This level of statistical control is important given the need to ensure findings can be meaningfully applied to the outcomes measured. An additional multiple regression statistical procedure was applied for the analysis related to Grade 3 MAP-R due to the lack of data related to prior performance. For this analysis, students who were indicated as having "rapid guessing[1]" were excluded. Effect sizes, measured by Cohen's *d*, also were computed to establish the magnitude of the performance differences associated with the assessment setting. For the purpose of this analysis *d*=0.2 to .49 represents a small effect size, *d*=0.5 to .79 represents a medium effect size, and *d=0.8* or higher represents a large effect size. Tables 7 to 12 display the

[1] A rapid guess is when a student's response is below the normal response time measured by NWEA and is so fast that the student could not have viewed the question completely. When a student guesses rapidly multiple times, the test automatically pauses and alerts the Proctor. A rapid-guessing threshold of 30% of questions gives leaders the opportunity to retest (NWEA, 2019).

raw means, adjusted mean difference, and corresponding effect size (if the difference is statistically significant, $p < .05$).

RIT Scores for MAP-M

Tables 7 to 9 provide a comparison of virtual and in-person mean MAP-M RIT scores by grade level, students receiving special education and those identified as LEP. When race/ethnicity, services received, and prior achievement were controlled, data suggested that for students in Grades 2 through 8 who took the test virtually, lower MAP-M RIT scores were observed compared to those taking the assessment in-person; however, the effects were small ($d =$.20 to .049) or not practically significant indicating the differences observed were not large enough to have a meaningful impact. Given these small effects, there is no significant evidence to suggest that test environment meaningfully impacts MAP-M test scores. Interestingly, Grade 1 students in total and Grade 1 who received special education services, and who identified as LEP all evidenced higher mean MAP-M RIT scores in the virtual testing setting. These data suggest students in lower grades performed better on the virtual administration of MAP in the area of math. However, the size of the effects remained small (d=0.2 to .49).

Table 7

Mean MAP-M Fall 2020 Compared to Fall 2019 RIT Scores for All Students by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (p)	Effect Size (*d*)
1	9,830	170.76	16.37	10,059	167.11	16.78	3.76	466.3	.000	.23
2	10,316	182.80	16.61	10,358	188.04	18.02	-5.34	1,291.0	.000	.30
3	10,343	191.51	15.50	10,656	193.23	14.10	-1.72	226.7	.000	.12
4	10,795	202.16	15.86	11,032	205.12	14.83	-3.34	1,071.0	.000	.22
5	11,117	213.18	19.11	11,393	216.41	17.28	-3.59	1,211.9	.000	.20
6	10,489	219.37	17.00	11,278	221.31	17.13	-2.14	512.6	.000	.13
7	10,567	227.72	18.68	11,251	228.05	18.87	-1.48	238.2	.000	.08
8	10,763	233.36	19.59	11,202	233.13	20.05	-1.16	140.2	.000	.06

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

Table 8

Mean MAP-M Fall 2020 Compared to Fall 2019 RIT Scores for Students Receiving Special Education Services by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (*p*)	Effect Size (*d*)
1	980	164.78	17.49	1,124	155.67	19.10	8.65	194.9	.000	.46
2	1,024	173.48	17.98	1,310	173.51	21.78			.517	
3	1,150	182.03	17.60	1,273	179.53	16.62	1.58	14.3	.000	.09
4	1,196	189.68	17.71	1,458	191.08	17.80	-2.01	26.4	.000	.11
5	1,340	197.06	19.95	1,391	198.54	19.58	-3.44	102.6	.000	.17
6	1,088	202.36	17.39	1,260	203.75	17.55	-1.44	17.8	.000	.08
7	1,040	209.07	18.59	1,255	209.44	18.86	-1.37	16.6	.000	.07
8	1,096	214.29	18.80	1,300	213.95	20.02	-1.21	13.5	.000	.06

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

Table 9

Mean MAP-M Fall 2020 Compared to Fall 2019 RIT Scores for Students Identified as LEP by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (*p*)	Effect Size (*d*)
1	2,689	164.85	17.63	3,075	158.78	15.32	7.32	431.2	.000	.44
2	3,117	175.69	17.41	3,213	179.31	17.24	-3.35	133.6	.000	.19
3	3,140	183.62	15.41	3,097	185.90	13.36	-1.48	39.7	.000	.10
4	3,109	192.97	14.78	3,064	195.91	13.76	-3.64	295.1	.000	.25
5	3,035	200.14	16.40	3,043	205.09	14.90	-4.83	527.4	.000	.30
6	2,452	206.95	14.37	2,388	207.28	13.65	-1.58	51.5	.000	.11
7	1,425	208.47	15.62	1,417	208.15	15.39	-1.67	31.6	.000	.11
8	1,080	210.91	17.30	1,026	209.86	17.42	-.93	6.3	.012	.05

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

RIT Scores for MAP-R

Results for MAP-R (Tables 10 through12) show that mean MAP-R RIT scores from the in-person administration were similar to scores from the virtual administration across grade levels for all students, students identified as LEP, and/or students receiving special education services. Although most differences were statistically significant across grade levels, the effect sizes were insignificant ($d < .20$) indicating the differences observed were not large enough to have a meaningful impact. One exception to this score pattern was observed in Grade 3 students receiving special education services where the adjusted mean MAP-R RIT Score was higher for the virtual administration compared to the in-person administration. The effect size was small ($d = .26$) indicating the difference observed was not large enough to be meaningful.

Table 10

MAP-R Fall 2020 Compared to Fall 2019 RIT Scores for All Students by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (p)	Effect Size (*d*)
3	11,166	191.13	19.22	11,836	191.21	19.23			.249	
4	10,848	201.62	18.27	11,045	202.75	17.70	-0.92	62.0	.000	.05
5	11,160	208.95	18.00	11,373	210.44	17.04	-1.01	89.6	.000	.06
6	10,417	216.13	16.77	11,518	216.75	16.33	-0.68	41.7	.000	.04
7	10,706	221.34	17.09	11,354	221.52	15.45	-0.48	21.0	.000	.03
8	10,870	225.31	17.00	11,203	225.38	15.67	-0.25	5.8	.016	.02

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

Table 11

MAP-R Fall 2020 Compared to Fall 2019 RIT Scores for Students Receiving Special Education Services by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (*p*)	Effect Size (*d*)
3	1,207	178.34	20.08	1,394	173.18	19.67	5.17	171.1	.000	.26
4	1,200	187.26	18.92	1,447	184.25	20.62	3.40	57.3	.000	.17
5	1,345	193.49	19.73	1,380	191.14	20.87	2.40	33.7	.000	.12
6	1,072	199.52	18.66	1,293	198.33	19.17	1.68	16.0	.000	.09
7	1,036	204.68	19.11	1,270	205.88	17.37			.772	
8	1,092	209.83	17.54	1,300	210.09	17.47			.226	

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

Table 12

Mean MAP-R Fall 2020 Compared to Fall 2019 RIT Scores for Students Identified as LEP Students by Grade Level

	Fall 2020 RIT (virtual)			Fall 2019 RIT (in-person)						
Grade	N	Mean	SD	N	Mean	SD	Adjusted Mean Difference	F	Sig (*p*)	Effect Size (*d*)
3	3,352	178.44	18.38	3,602	178.58	18.16			.281	
4	3,117	188.10	17.32	3,063	189.89	16.67	-1.26	25.7	.000	07
5	3,065	194.41	17.09	3,017	197.57	15.93	-1.68	47.4	.000	.10
6	2,434	201.80	15.33	2,510	201.89	15.01	-0.94	13.0	.000	.06
7	1,505	200.58	16.13	1,458	202.48	14.81	-2.19	36.2	.000	.14
8	1,130	202.03	17.13	1,033	203.47	15.82	-2.22	24.0	.000	.13

SD = Standard Deviation. Statistically Significant: $p < 0.05$.
Practically Significant: d=0.2 to .49 represents a small effect size, d=0.5 to .79 represents a medium effect size, and $d \geq 0.8$ represents a large effect size.

Findings

- Although there were differences in test duration in MAP assessments conducted in virtual versus in-person settings, most differences were small yielding an average of 5-minutes of difference in testing time. What this means, practically, is that students generally spent the same average amount of time on testing regardless of how the test was administered. This finding persisted despite the fact that there were noted technical issues that emerged during the virtual administration of MAP assessments.
- Student grade level also impacted the difference observed in test duration. With regard to MAP-R assessments, students in Grade 5 took longer to test in-person when compared to the virtual setting. However, the average difference in test duration was less than 10 minutes.
- The biggest differences in test duration were observed in MAP-R among students identified as LEP. In general, this group of students took less time to test in the virtual versus in-person setting. Differences varied across grade levels with the largest difference occurring in Grade 5.
- Results for MAP-R show that mean score differences associated with the assessment setting (in-person vs. virtual) were generally negligible. The data suggest that student MAP-R score performance is comparable regardless of the testing setting.
- In general, students who took MAP-M in the virtual setting took a longer time when compared to the in-person setting. However, the differences in test duration were noticeably small and varied among grades levels. Importantly, the smallest difference was observed in Grades 2, 4, and 5.
- There were notable differences in Grades 2, 4, and 5 students who showed lower MAP-M RIT scores in the virtual setting when compared scores from in-person setting. Similar results were observed for Grades 4 and 5 among students identified as LEP.
- Importantly, Grade 1 students receiving special services had higher scores in the virtual setting than in the in-person setting. Overall, the mostly small or insignificant effect sizes for MAP-M suggest that the performance differences are not be substantial enough to suggest that test setting made a considerable difference.

Additional Considerations

This analysis used fall MAP RIT scores; further analysis is needed to make sufficient conclusions about MAP Growth performance data beyond the fall 2020 administration and to examine whether the extended school-building closures during the 2020–2021 school year would have a more profound impact on student performance.

Reliability of MAP-Mathematics and MAP-Reading Test Scores: Fall 2016 to Fall 2020

Nyambura Maina, Ph.D.
Julie Wade, M.A.

Purpose and Data Sources

The change to a virtual test administration during the fall 2020 MAP assessment administration period introduced a number of questions about the reliability of MAP test scores. The current analysis sought to shed light on the reliability of MAP test data by examining the Standard Error of Measurement (SEM) for MAP-R/M tests administered from fall 2016 to fall 2020. Data from fall 2016 to fall 2019 were used to establish the performance pattern of students who tested in the in-person setting. Data from fall 2020 were used to represent student performance for students who tested in the virtual setting. The measure of reliability was determined by examining the SEM for the fall 2020 virtual administration and comparing the findings with the SEM for in-person MAP administrations from fall 2016 to fall 2019.

Definition of Terms

Standard Error of Measurement

On testing with MAP, the SEM represents the level of accuracy and precision in the student's test performance (Jensen, 2015). The smaller the SEM, the more precise the measurement capacity of the instrument. A higher SEM indicates less consistency in the student's pattern of responses and suggests it is not a reliable indicator of performance. According to data provided by NWEA, the typical observed standard errors on math and reading MAP tests range from approximately 2.9 to 3.5 (NWEA, 2020). Importantly, the MAP assessment is considered valid with an SEM up to 5.5. The SEM is related to the confidence interval of the score[2], (i.e., the range of scores that show, with some measure of confidence, where the "true" score lies). For example, if a student received an observed RIT score of 188 with an SEM of 3.00, the 95% confidence interval would be about 182 to 194 (1.96 x 3.0 = 5.9). If the SEM is larger, say 5.0, then the 95% confidence interval would be about 178 to 198 (1.96 x 5.0 = 9.8), indicating less precision in the range of possible scores for that student.

Analysis of Standard Error of Measurement

The SEM statistics were used to estimate the reliability of MAP-M and MAP-R test scores across the assessment periods. To compute SEM by grade and subgroup levels, yearly MAP reading and mathematics data files were merged with enrollment data to obtain the student demographic information. Descriptive statistics were used to examine the SEM observed for in-person (fall 2016 through fall 2019) and virtual (fall 2020) test administrations. The SEM for each student's MAP score is calculated and reported by NWEA in their score reports. For this analysis, the mean SEM were calculated for all students in each grade for each year examined. When SEM were examined for student subgroups, the mean SEM were calculated for each subgroup within each grade, by year of test administration.

[2] To compute the 95% confidence interval the SEM is multiplied by 1.96 to determine the range around the observed score in which a score would fall if the student were tested the next week or took a different version of the same test.

MAP-M Mean SEM across Five years (2016 to 2020)

Table 1 depicts the mean standard errors of MAP-M scores by grade level. Mean SEM in fall 2020 ranged from 2.98 to 3.32 across grade levels, and all mean SEM were within the "typical" range indicated by NWEA (2.9 to 3.5). The consistency of SEM across the five years of administrations, including fall 2020, when students took the assessments virtually, provides further evidence that the fall 2020 MAP Growth scores were, on average, a reliable estimate of students' achievement.

Table 1

Mean Standard Errors of MAP Mathematics by Grade

Grade		Fall 2020	Fall 2019	Fall 2018	Fall 2017	Fall 2016
K	Mean SE	3.32	3.26	3.25	3.25	2.95
	(N)	(9,682)	(11,243)	(10,892)	(10,803)	(10,947)
1	Mean SE	3.29	3.23	3.23	3.23	2.94
	(N)	(10,812)	(11,639)	(11,325)	(11,401)	(11,629)
2	Mean SE	3.27	3.24	3.23	3.22	2.94
	(N)	(11,039)	(11,622)	(11,574)	(11,758)	(11,813)
3	Mean SE	2.99	2.92	2.92	2.92	2.92
	(N)	(11,228)	(11,878)	(11,834)	(11,997)	(12,344)
4	Mean SE	2.98	2.92	2.92	2.92	2.93
	(N)	(11,468)	(12,096)	(12,187)	(12,438)	(12,029)
5	Mean SE	3.00	2.94	2.94	2.94	2.96
	(N)	(11,739)	(12,339)	(12,504)	(12,200)	(11,943)
6	Mean SE	2.99	2.93	2.94	2.94	2.97
	(N)	(11,109)	(12,250)	(12,117)	(11,873)	(11,126)
7	Mean SE	3.02	2.93	2.94	2.94	2.98
	(N)	(11,322)	(12,229)	(11,996)	(11,251)	(11,122)

SEM of MAP-Reading. Table 2 depicts the mean standard errors of MAP-R scores by grade level. The standard errors were similar across all grade levels. Mean SEM in fall 2020 ranged from 3.36 to 3.38 across grade levels, well within the range indicated as reliable by NWEA (2.9 to 3.5). Again, the mean SEM for reading were consistent over the years across all grade levels, and all mean SEM were within NWEA's "typical" range, providing evidence that the scores were, on average, reliable—or that the reliability of the test was not influenced by taking the test in class or a virtual setting.

Table 2

Mean Standard Errors of MAP Reading by Grade

Grade		Fall 2020	Fall 2019	Fall 2018	Fall 2017	Fall 2016
3	Mean SE	3.38	3.38	3.37	3.37	3.37
	(N)	(11,194)	(11,856)	(11,792)	(11,918)	(12,320)
4	Mean SE	3.36	3.37	3.36	3.36	3.36
	(N)	(11,501)	(12,090)	(12,122)	(12,402)	(11,983)
5	Mean SE	3.36	3.39	3.38	3.36	3.36
	(N)	(11,742)	(12,340)	(12,468)	(12,180)	(11,901)
6	Mean SE	3.36	3.36	3.37	3.35	3.36
	(N)	(10,992)	(12,500)	(12,102)	(11,864)	(11,552)
7	Mean SE	3.36	3.37	3.38	3.37	3.37
	(N)	(11,220)	(12,332)	(11,966)	(11,668)	(11,618)
8	Mean SE	3.37	3.39	3.40	3.38	3.38
	(N)	(11,343)	(12,063)	(11,984)	(11,783)	(11,367)

SEM for MAP-M and MAP-R for student groups

Mean SEM for students in specific student groups within each grade were examined for three years (2018, 2019, and 2020) of MAP-M and MAP-R assessments[3]. Student groups included students receiving special education services, students identified as Limited English Proficient (LEP), and students who received Free and Reduced-price Meals Systems (FARMS) services within MCPS focus groups. The standard errors were similar across the years for each student group within grade level regardless of whether the assessment was conducted virtually (fall 2020) or in-person (fall 2019 and fall 2018). Looking across all student groups by year, the SEM for mathematics assessments ranged from 2.98 to 3.30 in fall 2020 and from 2.91 to 3.28 in fall 2019. Similarly, the SEMs for the MAP-R assessments ranged from 3.35 to 3.41 in fall 2020, and from 3.35 to 3.42 in fall 2019. These data indicated that the mean SEM were consistent across the student groups and across years, and all mean SEM were within the "typical" range indicated by NWEA (2.9 to 3.5). These analyses provide evidence that the mathematics and reading scores for students in these groups were, on average, reliable and that the properties of the test were not influenced by taking the test in-person or virtually.

[3] Detailed tables with SEM for student groups within grades are available from the authors.

Findings

- Fall 2020 MAP scores in mathematics and reading were shown to be reliable estimates of students' performance. The mean Standard Error of Measurement (SEM) for each grade was well within the range considered reliable (NWEA, 2020).
- The SEM of the fall 2020 MAP, administered in the virtual setting, were consistent with the SEM of previous administrations that occurred in the in-person setting. The consistency of SEM across the years demonstrates that the reliability of the MAP scores did not change with the administration in a virtual setting in fall 2020 compared with in-person administration in previous years.

Additional Considerations

Examination of SEM provided a reference point for assessing the confidence level of fall 2020 MAP test administration. Examination of SEM provided a reference point for assessing the confidence level of fall 2020 MAP test administration. The reliability for the fall 2020 test administration are corroborated by a study examining marginal reliability and test-retest reliability virtual and in-person administrations of MAP in districts across the country (Kuhfeld et al., 2020). Taken together, these measures of reliability, consistent with measures of previous years, suggest that educators can have confidence in the fall 2020 MAP administration.

The findings showed that the MAP Growth scores accurately represented the student performance on the tests taken in the fall 2020 virtual setting. Still, it is worth noting that scores from MAP Growth tests are estimates of student performance on a given day. Therefore, no score should be treated as an absolute or used in isolation. If the performance level is not as expected, it is not necessarily an indication that MAP data are not reliable. Because the RIT scores should correlate highly with state scores and other assessments, further validating the students' performance levels by retesting and examining other indices of performance is essential. Since the reliability of the MAP Growth assessments has been established in these analyses, additional data (such as classroom performance), the time students took, and projected growth can be used to provide a more complete picture of student achievement. In addition, standard errors, even when examined alongside time taken, may not accurately reflect the amount of effort that a student gave on the test—further calling for the use of multiple measures to ascertain the precise level of student performance.

Future analyses may provide additional information about the reliability of MAP Growth assessments in different test settings. Since all MCPS students tested in fall 2020 took the MAP assessments virtually, no comparisons based on differences in non-virtual settings were possible. This limitation applies both to measures of achievement or growth, as well as to the assessment of test reliability as it relates to a variety of test-taking environments. Future test administrations may offer further opportunities to study the effects of virtual learning and virtual test-taking. It will be important to learn as much as we can about the wide-ranging impact of the current modifications in teaching, learning, and the settings for administering assessments.

Differences in the Conditional Growth Index (CGI) Between In-Person and Virtual Test Settings

Shihching Jessica Liu, M.A.
Heather Wilson, Ph.D.

Purpose and Data Source

In response to the recent flurry of national articles and research highlighting concerns around the comparability of MAP scores between the virtual and in-person test settings (Huff, 2020; Kuhfeld & Karasawa, 2020; Nodan, 2020), this analysis examines differences in MCPS MAP Growth scores between all 2018 and 2019 in-person test setting and fall 2020 virtual test setting. Doing so allows us to determine the comparability of MAP Growth scores across settings, and identifies how these data can be used to assess students' educational progress and determine instructional needs. The Conditional Growth Index (CGI) was selected as the outcome measure for this analysis because it allows us to average scores and compare differences in student performance between in-person and virtual test settings across grades.

Definition of Terms

What is the Conditional Growth Index?

The Conditional Growth Index (CGI) places students' performance on a standardized and comparable scale and expresses student growth relative to the growth of a national sample of peers that have been matched by grade, RIT score and weeks of instruction (Thum & Kuhfeld, 2020). This type of matching provides an opportunity for student growth to be compared across groups of students in various grades, subject areas and achievement levels (i.e., high, moderate, low achievers) (NWEA, 2021).

The CGI is calculated using the following equation:

$$observed\ RIT\ growth - projected\ RIT \frac{growth}{standard} deviation.$$

The resulting CGI score is expressed in standard deviation units. A CGI value of zero corresponds to the typical growth for similar students, indicating student growth exactly matched the projections. Said another way, a CGI score of zero tells us that a student demonstrated the same amount of growth compared to the student growth norms[4]. If converted to the commonly used metric in education, a percentile (as NWEA does with the Conditional Growth Percentile), a CGI of zero corresponds to a percentile rank of 50. This means the student's observed growth compared to their growth projection was greater than 50% of all students in the NWEA national norm group. The CGI is used rather than a percentile metric when averaging the growth of groups of students because students in the same percentile rank can have different growth and the difference in growth between two adjacent percentiles can be different ("Understanding CGI and CGP", 2020).

[4] Student growth norms are the 2020 NWEA student or school growth norms. These growth norms indicate median growth levels for students or schools based on their grade, starting RIT score, the subject in which they tested, and the amount of instructional time between two test events. (NWEA, 2021).

If a student demonstrated growth equivalent to the norm, their CGI score would be zero. This is the case regardless of student grade, subject, or starting achievement level. If the CGI score is positive it means the student's growth exceeded the growth norms, whereas if the CGI score is negative the student's growth was less than the growth norms. As an example, a CGI score of 1.0 means a student's growth is one standard deviation above the growth norm, whereas a CGI score of -1.0 means a student's growth is one standard deviation below the growth norm. A CGI greater than 1 would represent high level of growth and a CGI below – 1.0 would represent a substantial decrease in growth.

Analysis

To determine if there were differences in mean CGI scores between the in-person and virtual test setting for fall 2018 (in-person), fall 2019 (in-person) and fall 2020 (virtual administration), we used a one-way repeated measures ANOVA. Statistical significance was set at $p \leq 0.05$ level.

We examined both MAP-R and MAP-M. Given the calculation of the CGI score requires two years of growth data to calculate the difference between projected scores and observed scores, the MAP-R analysis includes students with CGI scores from Grade 6 to Grade 8. The analysis for MAP-M includes students with CGI scores from Grade 3 to Grade 9.

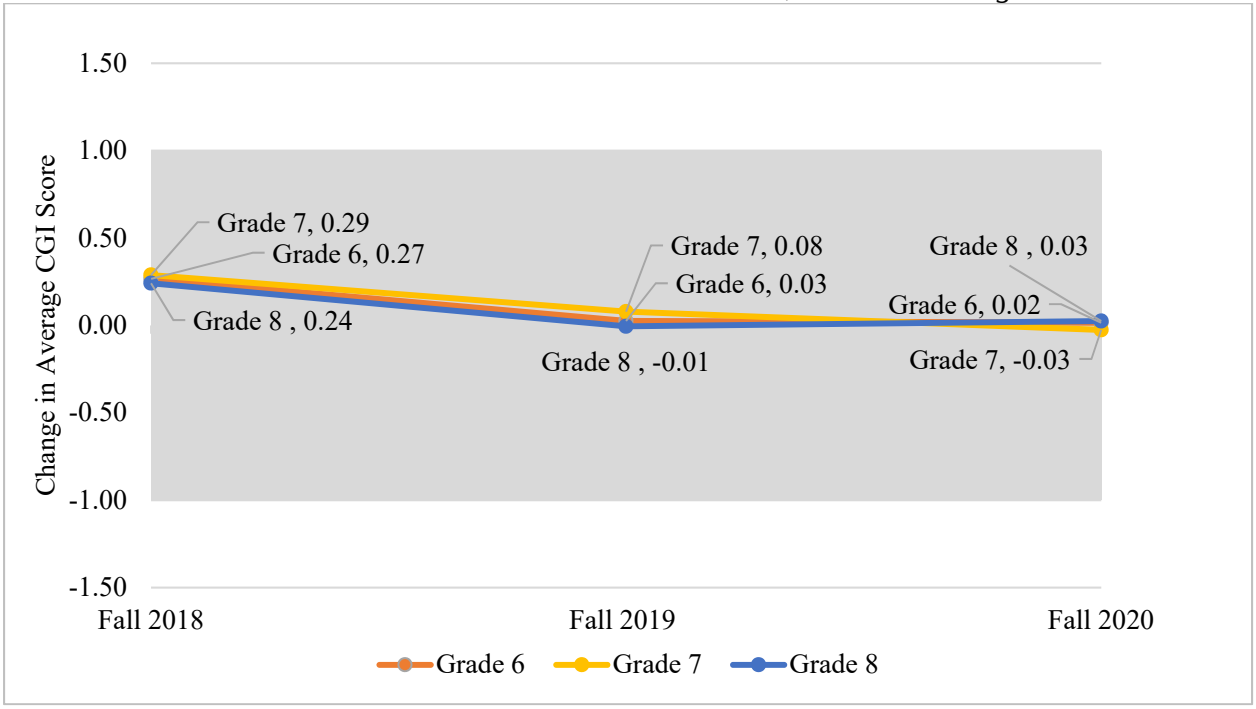
A key limitation of this analysis is that it tracks a group of students longitudinally rather than using a comparison group. All students included in the analysis experienced the same change in test setting from in-person to virtual during the fall 2020 MAP Growth administration. As such, any changes in student's CGI scores were not attributed to the change in test setting. Furthermore, students included in the current analysis all experienced the same testing and instructional environments (i.e., in-person vs virtual) during the years examined. The lack of MAP-R data for Grades 3 to 5 restricts the comparison of CGI scores between reading and mathematics.

MAP-R Mean CGI Scores

Figure 1 displays the mean CGI for MAP-R by grade across the three administrations. The grey shading represents the typical growth range (-1.0 to 1.0). Average student growth as measured by the CGI was within normal range, for each test administration. Further, when looking across test settings the average growth was lower in the virtual setting (fall 2020) compared to the in-person settings (fall 2018 and 2019). The average CGI for reading ranged from -0.03 to 0.29, a small amount of difference in scores and well within the expected range for student growth. In summary, average student reading growth, as measured using the CGI, was lower in the virtual setting (fall 2020) compared to the in-person settings (fall 2018 and 2019).

Figure 1

Mean MAP-R CGI Scores across Three Test Administrations, Grades 6 through Grade 8

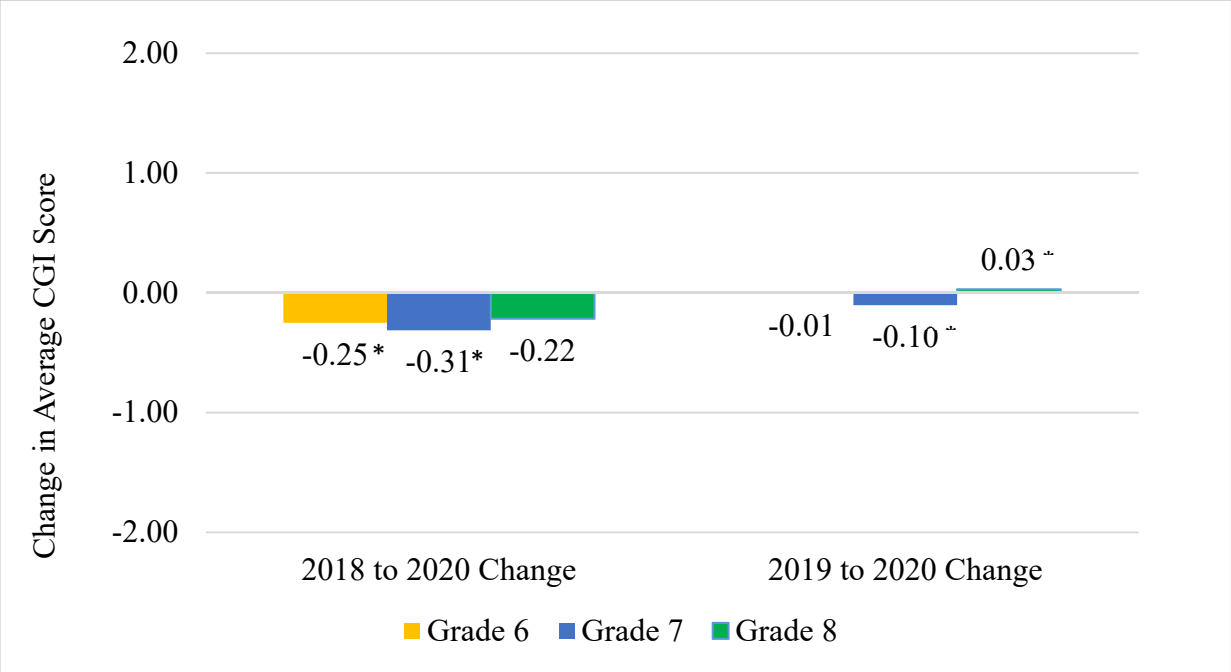


MAP-R: Average Change in Student Growth Scores Between Virtual and In-Person Test Settings

This analysis explored whether there was a statistical difference in mean CGI scores for students tested in the virtual and in-person settings. Figure 2 shows the average change in MAP-R CGI scores from the in-person test setting (fall 2018 and 2019) compared to the virtual test setting in fall 2020. Statistical significance was set at $p \leq 0.05$ level. The asterisks indicate results were statistically significant.

Figure 2

Average Change in MAP-Reading CGI Scores Between In-Person and Virtual Test Settings for Grades 6 through 8



Additional analysis was conducted to examine the amount of change that occurred in CGI scores from the in-person test setting in fall 2018 compared to the virtual test setting in fall 2020. Accordingly, there were small changes observed in CGI scores across Grades 6, 7, and 8 as shown in Figure 2. However, the changes fell within the normal range of change in growth. While there were statistically significant differences in CGI scores from the in-person (2018) to the virtual setting (2020) for Grades 6 and 7, the observed changes fell within the normal range of growth and there is not significant evidence to suggest that test environment meaningfully impacts MAP-R test scores.

The average change in CGI scores across Grades 6, 7, and 8 were small from fall 2019 to fall 2020 as well. Importantly, for Grades 6 and 7, the analysis showed student growth was slightly lower in the virtual test setting than in the in-person setting; however, the decreases in growth were relatively small (-0.01 for Grade 6, -0.10 for Grade 7). For Grade 8, a slight gain in reading CGI growth (0.03) for students that participated in the virtual test setting was observe. Changes in CGI for Grades 7 and 8 were statistically significant from 2019 to 2020 ($p \leq .05$). Overall, student growth changed very little from the in-person test setting to the virtual test setting from fall 2018 to fall 2020.

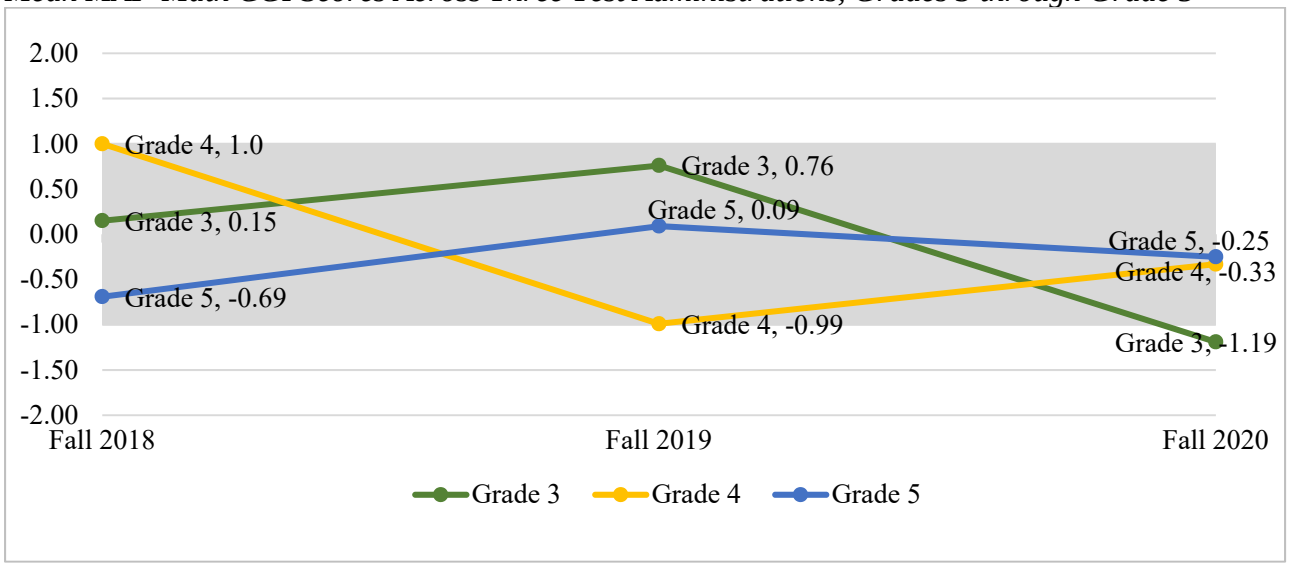
The average change in MAP-R growth, as measured by CGI, was lower in the virtual test setting when compared to the in-person setting (Figure 2). While the growth for both groups was within the expected growth index, students that tested in the virtual setting evidenced a smaller amount of growth when compared to students tested in the in-person environment. The only exception was Grade 8 from fall 2019 to fall 2020. The largest changes in average CGI scores occurred from fall 2018 to fall 2020, although the changes were small. While four of the changes were statistically significant, the average change in CGI scores between the in-person and virtual setting for reading were small ranging from -0.31 to 0.03. Given these findings, there is little evidence to suggest that the type of test setting significantly impacted MAP-R scores.

Elementary MAP-M Mean CGI Scores

Figure 3 displays the mean CGI for MAP-M by elementary grade across the three test administrations. The average CGI ranged from -1.19 to 1.0. Mean growth in math for Grades 3, 4, and 5 were within normal range (-1.0 to 1.0), for students that tested in the virtual setting and in-person setting. The one exception is Grade 3 in fall 2020 where a negative CGI of -1.19 was observed, which means student scores fell outside the range of normal growth. When looking across test settings, the mean CGI scores for only Grade 3 decreased from the in-person settings (fall 2018 or 2019) compared to the virtual setting (fall 2020). It could be that Grade 3 students were impacted more by the loss of instruction due to school-building closures because they were in Grade 2 at the time of the closing. Conversely, while the mean CGI for Grades 4 and 5 fluctuated across test administrations and were within the normal growth range, their performance was not necessarily lower in the virtual setting than the in-person setting.

Figure 3

Mean MAP-Math CGI Scores Across Three Test Administrations, Grades 3 through Grade 5

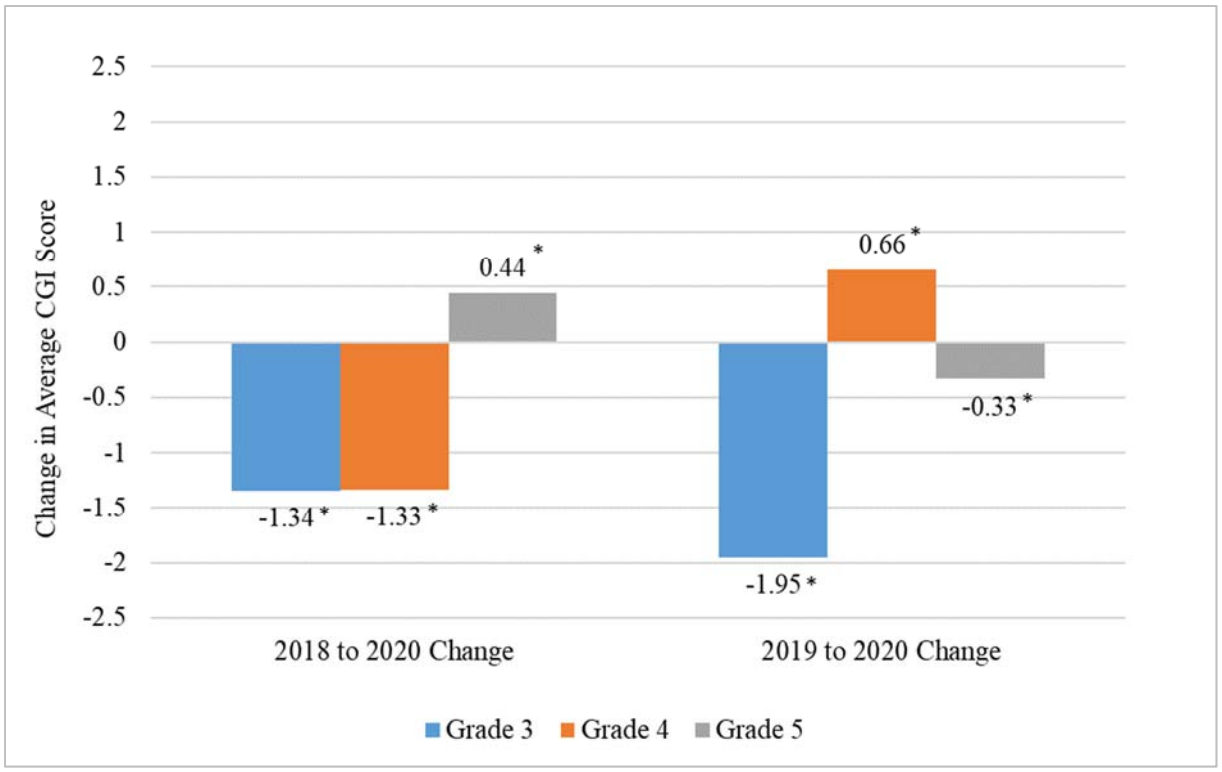


Elementary MAP-M: Mean Change in Student Growth Across Three Administrations

Figure 4 shows the average change in MAP-M CGI scores for Grades 3, 4, and 5 from the in-person test settings in fall 2018 and 2019 compared to the virtual test setting in fall 2020. Statistical significance was set at $p \leq 0.05$ level. The asterisks indicate results were statistically significant.

Figure 4

Average Change in MAP-M CGI Scores Between In-Person and Virtual Test Settings for Grades 3 through 5



Comparing the in-person test setting in fall 2018 with the virtual test setting in fall 2020, average change in math growth was mixed for Grades 3 through 5 (Figure 4). The CGI declined for Grades 3 (-1.34) and Grade 4 (-1.33) falling outside the normal growth ranges (-1.0 to 1.0). Alternatively, average math growth increased by 0.44 for Grade 5. All of the changes were statistically significant. Overall, between fall 2018 to fall 2020, student growth decreased substantially from the in-person setting to the virtual setting with the exception of Grade 5.

Similar to the fall 2018 to fall 2020 comparison, the average change in math growth was mixed for Grades 3 through 5 from fall 2019 to fall 2020. Notably, the decline of the Grade 3 CGI for this time period was -1.95, considerably below the normal growth range. For Grade 4 average

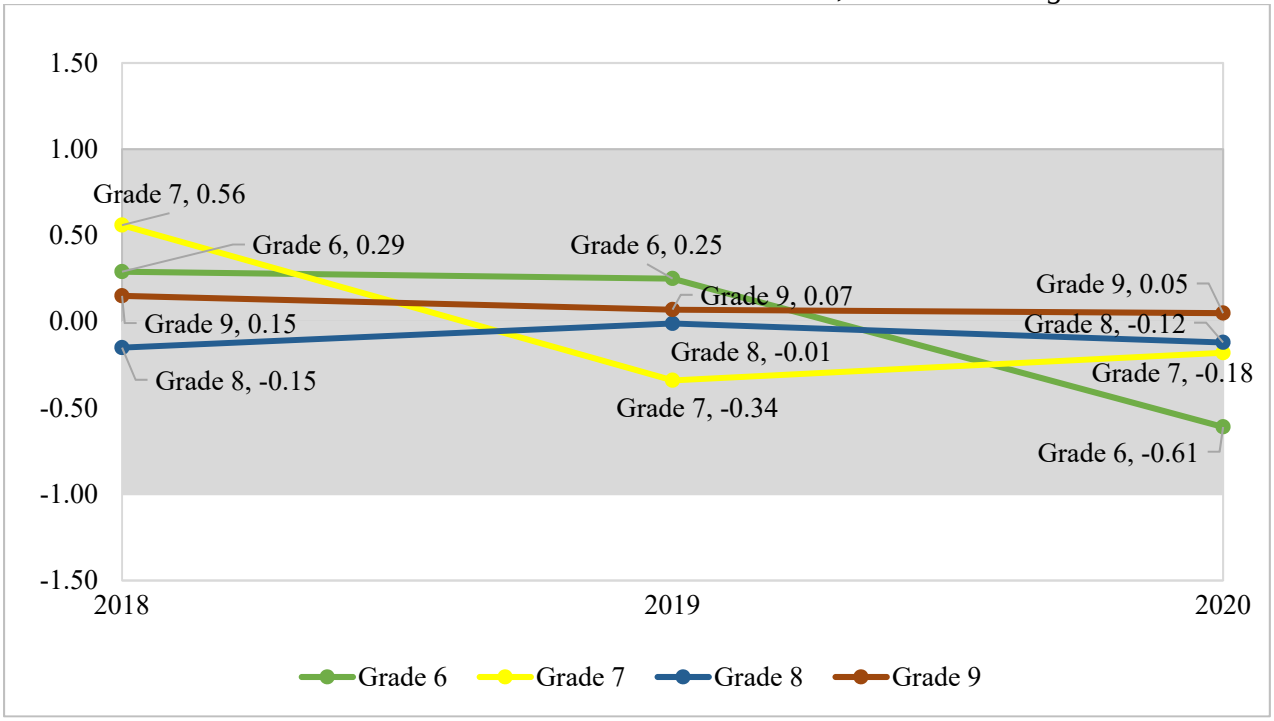
math growth increased by 0.66. For Grade 5, the average math growth decreased by -0.33. All of the changes were statistically significant. Overall, between fall 2019 to fall 2020, student growth decreased from the in-person setting to the virtual setting with the exception of Grade 4.

Secondary MAP-M: Mean CGI Scores for Fall 2018, Fall 2019, and Fall 2020

Figure 5 displays the mean CGI scores for Grades 6 through 9 across the three test administrations. The average CGI ranged from -0.61 to 0.56 within the normal growth range. Further, when looking across test settings, the average change in student growth for Grades 6 and 7 were lower in the virtual setting (fall 2020) compared to the in-person settings (fall 2018 or 2019). Whereas, the 2020 average growth for Grades 8 and 9[5] in the virtual setting was similar to the in-person setting.

Figure 5

Mean MAP-Math CGI Scores Across Three Test Administrations, Grades 6 through and Grade 9



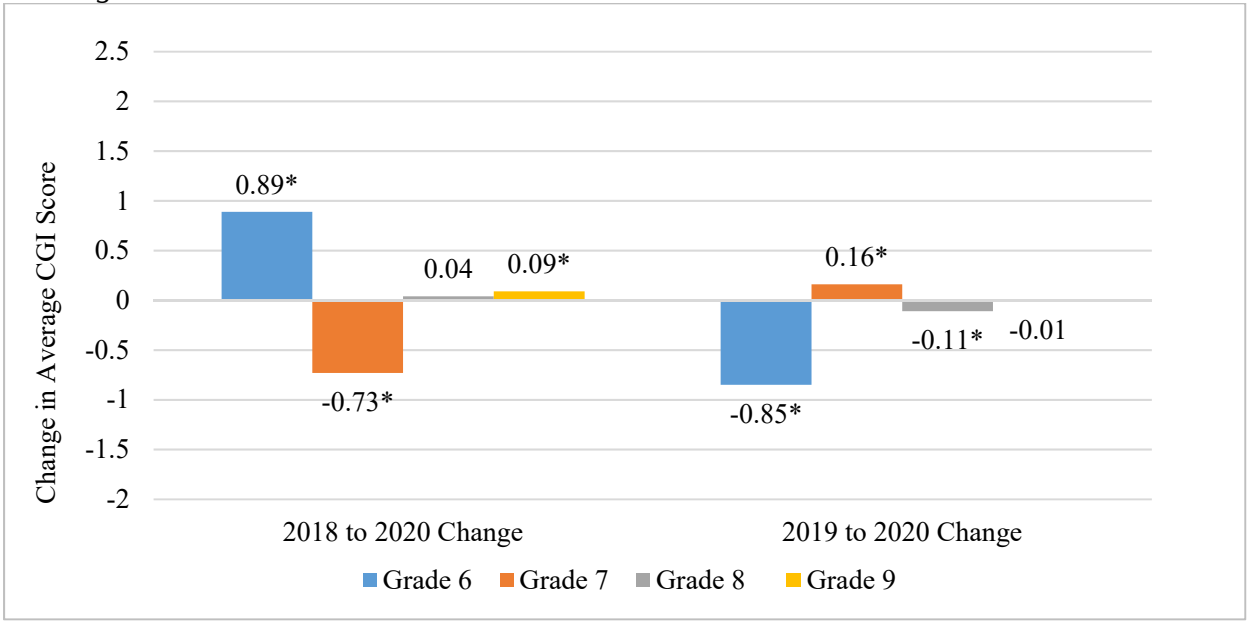
Secondary MAP-M: Mean Change in Student Growth Across Three Administrations

Figure 6 shows the average change in MAP-M CGI scores for Grades 6 through 9 from the in-person test settings in fall 2018 and 2019 compared to the virtual test setting in fall 2020. The asterisks indicate results were statistically significant at $p \leq 0.05$ level.

[5] It is important to note that Fall 2020 was the first year that high school students participated in MAP testing. Since MAP growth scores were created based on the MAP scores of previous two years, this cohort of Grade 9 students also took MAP tests in Grade 7 and Grade 8.

Figure 6

Average Change in MAP-M CGI Scores Between In-Person and Virtual Test Settings for Grades 6 through 9



Comparing the in-person test setting in fall 2018 with the virtual test setting in fall 2020, the average change in MAP M growth was mixed. (Figure 6). However, all average change in growth fell within normal growth range (-1.0 to 1.0). Grades 6 and 7 saw larger changes in student growth (0.89 and -0.73, respectively) in the virtual test setting than the in-person test setting. Alternatively, Grades 8 and 9 saw very small positive changes in student growth (0.04 and 0.09). All of the changes were statistically significant except for Grade 8 (0.04). Overall, between fall 2018 to fall 2020, student growth increased from the in-person setting to the virtual setting with the exception of Grade 7.

Comparing the in-person administration in fall 2019 with the virtual administration in fall 2020, the average change in math growth declined for most grades (Figure 6). Although declines were observed for most grades, it is important to note that the average change fell within the normal growth range (-1.0 to 1.0). Grade 6 saw the largest decline with an -0.85 average change in student growth. The changes in growth were small for Grade 7 (0.16), Grade 8 (-0.11), and Grade 9 (-0.01). All of the changes were statistically significant except for Grade 9 (-0.01). Given these findings, there is little evidence to suggest that the type of test setting significantly impacted MAP-M scores.

Findings

- This analysis indicates reading and math growth scores are comparable across modes of test administration (in-person or virtual). The mean CGI scores for both reading and math were within the normal range of growth regardless of test setting. However, there were grade-level differences in the amount of student growth on the math assessments, most notably for students in Grades 3 and 6.
- The average change in MAP-R growth, as measured by CGI, was lower in the virtual test setting when compared to the in-person setting. However, the change was small and generally fell within the range of expected normal growth. Grade-level differences in CGI scores were identified for students in Grades 6, 7, and 8 on the MAP-R assessment. However, the size of the difference was small and all the differences were within the normal growth range.
- Comparing the virtual test setting to the in-person test setting for MAP-M student growth from Grades 3 to 9 fluctuated; but, stayed within the growth norm range of -1.0 and 1.0. One exception to this growth pattern was observed in the CGI scores for students in Grade 3 on the MAP-M assessment where the CGI scores decreased and fell below the expected level of growth. Given this pattern was not uniformly observed across grade levels, it is hard to ascertain why this pattern of performance emerged. What is important note is that students across most grade levels demonstrated the expected level of growth regardless of setting in which they took the test.
- The decreases in MAP-M CGI scores for the elementary grades were greater than the average declines in mean MAP-M CGI scores for the secondary grades. These data indicate that students in lower grades are likely more sensitive to shifts in test settings when compared to students at higher grade levels, particularly in the area of mathematics. It is important to note, the CGI scores for MAP-R did not reveal a similar pattern because data were not available for students in the elementary grades.

Additional Consideration

This analysis shows that the average student growth for math and reading decreased from the in-person setting to the virtual setting. While the decrease was small, and within expected ranges of growth, the data indicates that MAP scores are sensitive enough to respond to the subtle changes that might occur when there is a change in test setting. Further research for the Grade 3 MAP-M scores could help us understand the differences in student growth observed across test administrations. In addition, research on student growth from the upcoming 2021 spring test administrations may provide more information on the comparability of student performance across test settings.

References

Huff, K. (2020). *New data from curriculum associates quantifies impact of COVID learning loss raises questions about at-home testing.* Research Brief. Curriculum Associates North Billerica, MA. https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/iready-diagnostic-results-understanding-student-needs-paper-2020.pdf.

Jensen, N. (2015). Making Sense of Standard Error of Measurement. https://www.nwea.org/blog/2015/making-sense-of-standard-error-of-measuremen.t

Kuhfeld, M., Tarasawa, B., Johnson, A., Ruzek, E. & Lewis, K. (2020a). *Learning during COVID-19: Initial findings on students' reading and math achievement and growth.* Research Brief. NWEA. https://www.ewa.org/sites/main/files/file-attachments/learning_during_covid-19_brief_nwea_nov2020_final.pdf?1606835922

Kuhfeld, M., Lewis, K, Meyer, P., & Tarasawa, B. (2020b). *Comparability analysis of remote and in-person MAP Growth testing in fall 2020*. Technical Brief. NWEA

Nodan, M. (2020) *Fall assessments to gauge 'COVID slide' may be skewed. Can districts use them?* Online article. K-12 Dive. https://www.newsbreak.com/news/2115389484485/fall-assessments-to-gauge-covid-slide-may-be-skewed-can-school-districts-use-them

NWEA (2013). *Student effort and test score accuracy.* Educational Blog. (https://www.nwea.org/blog/category/assessment-basics/).

NWEA (2021). *Understanding CGI and CGP*. https://nwea.force.com/nweaconnection/s/global-search/Conditional?language=en_US.