TECHNICAL BRIEF

Comparability analysis of remote and in-person MAP Growth testing in fall 2020

November 2020

Megan Kuhfeld, Karyn Lewis, Patrick Meyer, and Beth Tarasawa

Suggested citation: Kuhfeld, M., Lewis, K, Meyer, P., & Tarasawa, B. (2020). Comparability analysis of remote and in-person MAP Growth testing in fall 2020. NWEA.

Table of Contents

List of Tables

List of Figures

1. Executive Summary

This study examined the psychometric characteristics and indicators of test quality of MAP Growth tests that were administered remotely and in-person in fall 2020. Using test scores from over 535,000 K-8 students in 147 school districts (92 operating fully remotely this fall, 55 offering in-person instruction to all students), this study provides insight into the comparability of remote versus in-school assessment. We found high levels of marginal reliability and test engagement across all grades, as well as consistent trends in test scores for remote and in-person tests for students in grades 3-8. Taken together, these findings increase confidence in the quality of data gathered from remotely administered MAP Growth assessments in grades 3 and up.

Key findings were:

1. Marginal reliability was high (≥0.90) across all grades and subjects across both remote and in-person test administrations.

2. Between-term correlations were high (>0.70) across grades and subjects, regardless of testing modality, with the exception of students in first and second grade in fall 2020.

3. Test engagement and test duration between fall 2019 and fall 2020 were similar between remote and in-person test takers. Students' test engagement remained high both for students who tested remotely and in-person in fall 2020 across grades and subjects.

4. When comparing test duration between fall 2019 and fall 2020, moderately larger increases were observed for students who tested remotely in fall 2020 relative to students who tested in-person.

5. In grades 3 through 8, achievement percentiles stayed the same or dropped from fall 2019 to fall 2020, with trends similar for remote and in-person testers and larger percentile score drops in math than in reading.

6. Students who tested remotely in grades 1 and 2 grade in fall 2020 showed large improvements in their percentile rank since fall 2019; while in-person testers in grades 1 and 2 showed patterns more consistent with older students (percentiles stayed the same or dropped).

2. Introduction

In the NWEA fall 2020 COVID research studies, we present a series of analyses based on MAP® Growth™ data from the fall of the 2020-21 school year as well as prior academic years. A key assumption underlying the interpretation of these data is that the mode of assessment has little to no impact on test scores. However, there are concerns around remotely

administered assessments (e.g., increased distractions, unfamiliar virtual meeting software, potential connectivity challenges, among others) that call into question whether assessments that were administered remotely in fall 2020 can be considered comparable to assessments administered in person.

NWEA launched a program of research to probe the comparability of remote and in-person tests in spring of 2020 when the pandemic first forced schools to close and resort to virtual instruction and assessment. Our initial findings from this research, conducted with a subset of schools that tested remotely in spring 2020, provided encouraging evidence that remote and in-person tests showed comparability in psychometric characteristics as well as student test engagement. [i, ii] Specifically, the spring 2020 comparison found that less than one percent of items showed differential item functioning (DIF) by testing modality (less than the percentage expected by chance alone), and that remote testers showed similar levels of test engagement as students who tested in-person. This research brief updates and builds on those promising findings using fall 2020 data from a large sample of schools across the nation to further investigate the validity and comparability of remote assessments. By triangulating across a range of assessment characteristics including psychometric properties as well as indicators of test engagement, this brief sheds further light on the comparability of remote versus in-person assessment.

Specifically, we explored the following research questions:

1. Did the mode of administration (in-person versus remote) have any impact on the psychometric properties (specifically, marginal reliabilities and test-retest correlations) of the MAP Growth assessments?
2. Were changes in test duration and test engagement between the 2019-20 and 2020-21 school year similar between remote and in-person test takers?
3. Did remote testers in fall 2020 show significantly better test performance relative to in-person testers after adjusting for prior achievement and student/district characteristics? Did remote/in-person differences vary across subjects/grades/racial groups?

The first research question examined a primary concern that the assessment itself functions differently when administered in different assessment modalities. To answer this research question, we examined the reliability of the test when administered remotely or in person. If we can establish test reliability is consistent across remote and in-person settings, we may still expect differences in student performance if aspects of students' testing environment impact their motivation and ability to pay attention during the test. The second research question addressed this concern by examining indicators of student test effort across in-person and remote test settings.

Finally, a remaining question when comparing remote and in-person test performance outcomes is that any differences we may see may not be due to testing modality, but instead attributable to confounding differences between districts that opened in-person versus fully remote. Specifically, it is possible that these districts serve different student bodies, and it is these demographic differences, not testing modality, that drive any differences in performance across settings. We probed this possibility in our third research question by controlling for students' past performance when examining their fall 2020 test scores. Additionally, we examined within-group differences (e.g., comparing White students' performance in remote settings with the

White students who tested in-person this year) controlling for a set of school district characteristics to attempt to better isolate remote/in-person mode effects.

3. Data

The data for this study are from the NWEA anonymized longitudinal student achievement database. School districts use NWEA MAP Growth assessments to monitor elementary and secondary students' achievement and growth throughout the school year, with assessments typically administered in the fall, winter, and spring. We used the reading and math test scores of over 535,000 students, from kindergarten through eighth grade in 2,074 schools from across the United States across three time points: fall 2019, winter 2020, and fall 2020.

3.1. Longitudinal Sample Description

In this study, we followed multiple intact cohorts of students across the 2019-20 and 2020-21 school years. For example, one cohort of students started kindergarten in fall 2019 and entered first grade in fall 2020. The primary advantage of using an intact cohort is that we could compare each student's fall 2020 test performance to his or her own prior test score. A disadvantage is that students may have systematically dropped out of our sample this fall due to the disruptions of COVID-19. For more details on the attrition patterns in the MAP Growth data in fall 2020, see the attrition report.[iii] We separately examined every two-year grade pair from grades K-1 to grades 7-8.

Our sample consisted of a subset of schools and districts who tested with MAP Growth assessments where either (a) the district was operating fully remotely by the time testing occurred this fall, or (b) all students in districts had the option for in-person instruction this fall. NWEA does not currently have a student-level indicator of whether a student tested remotely or in-person in fall 2020. Therefore, we used an indicator of district reopening status (collected by *Education Week*[iv] for over 900 districts in the country) as a best proxy for the likelihood testing was administered remotely or in-person in fall 2020 (districts that had a hybrid reopening plan were excluded). Students who attended schools with remote learning only and no in-person instruction available were defined as "Remote testers." Students who attended schools with full-time, in-person instruction available for all students were defined as "In-person testers." However, it is likely that this classification is imperfect as some students in districts in which in-person instruction was available for all students still may have opted to learn and test remotely this fall. NWEA is developing an indicator to more precisely capture whether a test was administered remotely or in-person which will make it possible to compare data quality across testing modalities more systematically in future research.

3.2. Sample Descriptive Statistics

In total, our sample contained 535,000 students from 147 unique districts (55 remote, 92 in-person). Descriptive statistics of the sample suggested in-person and remote districts were demographically and geographically different from each other (see Table 1). Eighty-four percent of school districts in our sample that opened remotely were in urban or suburban areas, while only 31% of in-person districts were in urban/suburban areas. The average enrollment in

districts that opened remotely in fall 2020 was far larger than the districts that opened in-person. Overall, the sample size per grade ranged from 40,000 to 90,000 students, and the majority of students in the districts that opened in-person were White, while the students in the remote only districts were more racially diverse (see Table 2).

3.3. Data Quality Measures

Measures of achievement. We used student test scores from NWEA MAP Growth reading and math assessments in this study. MAP Growth is a computer adaptive test—which means the level of precision is similar for students across the achievement distribution, even for students above or below grade level —and is vertically scaled to allow for the estimation of gains across time. Test scores are reported on the RIT (Rasch unIT) scale, which is a linear transformation of the logit scale units from the Rasch item response theory model. In this study, we used both students' RIT scores and percentile scores calculated using the NWEA 2020 MAP Growth norms.[v]

Measures of test effort. We presume that remote testing takes place in a less controlled environment than in schools, given potential additional distractions and concerns about students receiving assistance from family or use of outside resources on the assessment. The potential for a qualitatively different testing experience in remote settings compared to in school raises important questions about the quality of data from remote testing. Given the additional challenges of testing in a home environment, an important indicator of data quality is whether students were able to stay engaged during a test. Test disengagement, specifically rapid-guessing—when a student answers a test question so quickly that they could not have understood its content and provided an effortful response—poses a substantial threat to test validity.[vi]

While the remote testing environment differs from in-school testing, the MAP Growth assessment includes features intended to identify rapid-guessing behaviors and provide information to students and proctors to encourage students to re-engage with the assessment. When MAP Growth assessments are administered in schools, a proctor in the testing room gives students a password and instructions on how to access the test, answers student questions during testing, and monitors student progress on a computer that displays each student's progress. In remote testing, proctors are not physically present with students and cannot visually monitor the students' testing environments. Instead, proctors and students communicated during remote testing using a variety of methods, including text messages, phone conversations, and online video conferencing software. When video conferencing was used, the proctors had a webcam view of all students being testing but could not actively monitor a student's test-taking environment. Regardless of where the assessment is administered, MAP Growth uses an "auto-pause" feature to identify rapid-guessing and address test-taking disengagement in real-time: after a pre-specified number of rapid guesses, the test is automatically paused, and a message is displayed on the student's computer screen informing them that they are rushing through the test and asking them to slow down. The test proctor also receives a notification of the auto-pause and must enter a passcode to resume the student's test, presumably after encouraging the student to answer questions effortfully. If rapid-guessing continues, the auto-pause feature may engage up to two additional times during the assessment.

Prior NWEA comparisons of test engagement for students who tested remotely versus in person provided encouraging evidence that test engagement was similar in both remote and in-school test administrations in spring 2020.[ii] We expanded on this initial promising finding with more recent data by examining whether student engagement on MAP Growth assessments differs in remote testing on two measures of student test taking engagement. First, we examined trends in students' Response Time Effort (RTE), which indicates the proportion of responses that were solution behaviors rather than rapid guesses.[vii] Second, we looked at changes in overall test duration (measured as the number of minutes elapsed between the start and end of the test, excluding any pauses) between fall 2019 and fall 2020.

4. Methods

4.1. **Psychometric Properties**

We calculated the marginal reliability (for more information on the calculation of reliability, see NWEA, 2019[viii]) and test-retest reliability. We calculated these metrics for three testing terms (fall 2019, winter 2020, and fall 2020). We specifically included two pre-COVID-19 terms to establish a baseline against which to evaluate fall 2020 metrics.

4.2. Test Effort

We examined whether changes in student test effort between fall 2019 and fall 2020 varied dependent on whether students tested remotely or in-person. Specifically, we calculated RTE and average test duration by grade, term, subject, and plotted trends within each group separately by fall 2020 testing modality.

4.3. **Test Performance**

As noted above, simple between-group comparisons of test performance for in-person testers and remote testers may be uninformative or even misleading because of potentially confounding differences between districts that opened for in-person instruction and those that remained remote. To address this, we examined within-student changes in percentile rank between fall 2019 and fall 2020 separately by grade, subject, and fall 2020 testing modality. Within-student comparisons allow us to account for potential baseline differences in achievement between students in districts that opened remotely or in-person this fall. If testing remotely is associated with systematic changes in student achievement, we would expect to see different patterns in shifts in percentile rank over time between the two testing modalities.

In addition, to further rule out the explanation that differences observed between remote and in-person test scores in fall 2020 could be due to pre-existing differences between students in the two sets of districts, we ran a series of regression models that tested for differences in scores by testing modality while controlling for students' prior achievement, student demographic characteristics, and district characteristics. Specifically, we regressed the fall 2020 RIT scores on the fall 2019 RIT scores, an indicator for testing modality (where in-person is the reference group) in fall 2020, and a set of student- (indicators of race/ethnicity) and district-level covariates (district SES, percentage of English Learner (EL) students, percentage of special education students, and urbanity of district). We also thought it possible that the impact of testing modality

could differ across student groups, if, for instance, some student groups are more likely than others to experience home environments that are less conducive to ideal testing conditions. To allow for potential differences across student groups in the impact of testing remotely, we included a set of interaction terms between student race/ethnicity and attending school remotely in fall 2020. In this model, we estimated cluster robust standard errors to account for nesting of students in districts.

$$\begin{aligned}\text{Fall20score}_\text{i} = \ & \beta_0 + \beta_1\text{Fall19score}_\text{i} + \beta_2\text{RemoteF20}_\text{i} + \beta_3(\text{Fall19score}_\text{i} * \text{RemoteF20}_\text{i}) + \\ & \beta_4\text{Black}_\text{i} + \beta_5\text{Asian}_\text{i} + \beta_6\text{Hispanic}_\text{i} + \beta_7\text{OtherRace}_\text{i} + \\ & \beta_8(\text{Black}_\text{i} * \text{RemoteF20}_\text{i}) + \beta_9(\text{Asian}_\text{i} * \text{RemoteF20}_\text{i}) + \\ & \beta_{10}(\text{Hispanic}_\text{i} * \text{RemoteF20}_\text{i}) + \beta_{11}(\text{OtherRace}_\text{i} * \text{RemoteF20}_\text{i}) + \cdots + e_i.\end{aligned}$$

From this model, we estimated within-group differences in fall 2020 RIT score between remote testers and in-person testers (controlling for prior test score and district demographic characteristics) separately for each racial/ethnic group:

$$\begin{aligned}&\text{White Gap: } \hat{\beta}_2 \\ &\text{Black Gap: } \hat{\beta}_2 + \hat{\beta}_8 \\ &\text{Asian Gap: } \hat{\beta}_2 + \hat{\beta}_9 \\ &\text{Hispanic Gap: } \hat{\beta}_2 + \hat{\beta}_{10}\end{aligned}$$

5. Results

5.1. Psychometric Properties

We examined whether the assessment modality was associated with the reliability and across-time stability of students' test scores. Marginal reliability was high (≥0.90) across all grades and subjects across both remote and in-person test administrations (see Table 3). Between-term correlations were generally high (>0.70) across grades, subjects, and reopening status (see Table 4) with only few exceptions. For students in first and second grades who tested remotely in fall 2020, we saw lower test-retest correlations between pre-COVID-19 test scores and fall 2020 test scores; the same was not true of first- and second-grade students who tested in person. This suggests that test scores were less consistent from one testing period to the next for the youngest students who tested remotely, but not those who tested in person. Taken together, consistently high marginal reliabilities and test-retest correlations suggest that mode of administration appeared to have no adverse effects on the psychometric properties of MAP Growth, though differences were observed across test administration mode for students in the lowest grades.

5.2. Test Effort

Next, we examined if changes in test engagement and test duration between fall 2019 and fall 2020 were similar between remote and in-person test takers. Students' average RTE remained high across grades both for students who tested remotely and in-person in fall 2020 across grades (see Figure 1), indicating that testing remotely was not tied to a large decrease in test engagement this fall. Moderately larger increases in test score duration were observed for students who tested remotely in fall 2020 relative to students who tested in person, particularly

in the younger grades (see Figure 2). Although not displayed in Figures 1 and 2, average RTE and test duration in math were consistent with those depicted for reading.

5.3. Test Performance

We first conducted a descriptive analysis to examine whether performance shifted over time in different ways for remote testers compared to in-person testers. To do this, we plotted changes in students' median achievement percentiles from fall 2019 to fall 2020 separately by testing modality. As shown in Figures 3 (math) and 4 (reading), we found different patterns depending on grade level. In grades 3 through 8, test score percentiles stayed the same or dropped from fall 2019 to fall 2020, and these trends were similar for remote and in-person testers. Consistent with our research on learning loss across the COVID-19 period, we observe percentile rank drops were larger in math than in reading.[ix] However, in the early grades, we saw that trends in achievement shifts between fall 2019 and fall 2020 looked very different between remote testers and in-person testers. Specifically, students who tested remotely in first and second grade in fall 2020 showed large increases in their percentile rank since fall 2019; in contrast, in-person first- and second-grade testers showed patterns more consistent with older students (percentiles stayed the same or dropped). These results suggest that the remote testing experience is consistent with in-person testing for students in grade 3 and up but may qualitatively differ for the youngest students.

In addition to the descriptive analyses, we used regression models to explore if remote testers in fall 2020 showed significantly different test performance in fall 2020 relative to in-person testers after adjusting for prior achievement and student/district characteristics (see Table 5 for regression coefficients). In math, students who tested remotely typically scored higher than students who tested in person as indicated by a significant regression coefficient for our indicator of testing modality across grades. Although statistically significant, the remote testing advantage in math was slight (roughly 1 to 2 RIT points) for students in grade 3 and up and may have little practical significance. However, consistent with the findings from our descriptive analyses, remote testing advantages were especially notable for students in grades 1 and 2 in fall 2020, for whom the differences between in-person and remote testers were sizable in both math and reading (roughly 5 RIT points or more, which corresponds to over 0.25 SD). There was little evidence of a remote testing advantage in reading for older students who typically scored similarly or slightly lower compared with students who tested in person.

Finally, we also examined whether fall 2020 scores for remote versus in-person testers were different across subjects, grades, and racial/ethnic groups to understand whether students were differentially impacted by remote testing. The results of these within-group comparisons are plotted in Figure 5 (math) and Figure 6 (reading). We found that across grade levels, and relative to their same-race counterparts, Asian students showed the largest advantage from remote testing. Hispanic students were the most likely to perform lower on average when testing remotely, compared to Hispanic students who tested in person. This pattern was particularly evident in reading, though results varied somewhat by grade level.

6. Conclusion

Our analyses of the psychometric properties of the MAP Growth assessments provided general support for the comparability of scores from the two modes of testing for students in grade 3 and up. The results also showed similar test taking engagement, as measured by RTE, for remote and in-person assessments. This is encouraging, since rapid-guessing test disengagement poses a validity threat, and it seemed plausible that students could be less engaged without a proctor in the room with them. The high reliability for in-person and remote assessments, high test-retest correlations, and the similar trends in test scores for remote and in-person tests for students in grades 3-8 likewise increase confidence in the quality of data from remote assessments.

However, our results also raise questions that require additional exploration related to remote test data for students in K-2. In these grades, we found marginal reliability was high regardless of testing modality and test disengagement was low. We did however observe significantly lower test-retest correlations between in-person pre-COVID-19 test scores and scores from tests administered remotely in fall 2020 for students in these grades. We also observed longer average test durations and large increases in achievement percentiles for K-2 remote testers (but not in-person testers in these grades). Taken together, these findings suggest that remote testing may be a qualitatively different experience for the youngest students. Our data cannot speak to why we see different results for students in these grades and further research is needed to better understand these differences. Additional guidance and support may be needed to help schools and families establish testing conditions at home that are structured to be as similar as possible to in-person testing conditions, most especially for students in these earlier grades.

Finally, these analyses identified moderate differences in student performance between remote and in-person test administrations across race/ethnicity subgroups. Specifically, test scores of students from certain racial/ethnic groups were higher in remote testing conditions in comparison to their same-race peers who tested in-person, particularly in the earliest grades. Additional research can help us understand to what extent differences in home learning environments and economic and public health factors may be contributing to these differences across student groups.

In summary, the findings of these analyses strengthen confidence in the quality of the data from remote tests across most grades, with largely consistent findings in grade 3 and up across testing modalities. However, our results also indicate that caution is warranted when interpreting the test results for certain subsets of students who tested remotely, especially students in the earlier grades, and underscores the need for additional steps to be taken to ensure consistency in administration procedures for tests administered remotely in subsequent testing terms.

Table 1. Comparison of School Districts with Known Fall 2020 Reopening Plans That Tested in Fall 2019 and Fall 2020

	Full in-person reopening available for all students	Remote learning only
Average socioeconomic status (Standardized)	0.29	0.14
Average enrollment	7,102	28,351
Proportion urban	0.18	0.41
Proportion suburban	0.13	0.43
Proportion town	0.35	0.04
Proportion rural	0.35	0.11
Average % households with BA degree	0.24	0.34
Average % SPED	0.14	0.13
Average % ELL	0.04	0.12
Average % households in poverty	0.16	0.20
Average % SNAP receipt	0.10	0.11
Number of Districts	55	92

Note: SPED=Special Education, BA=bachelor's, ELL=English Language Learner. District demographic data comes from the Stanford Education Data Archive (SEDA).

Table 2. Sample Demographic Characteristics by Grade for Overall Sample and Broken Down by Fall 2020 Reopening Status

Grade	Male	White	Black	Other Race	Hispanic	Asian	Sample Size		
							Students	Schools	Districts
Overall Sample									
1	0.51	0.37	0.18	0.11	0.31	0.05	44,711	972	100
2	0.51	0.39	0.18	0.11	0.30	0.05	68,546	1,160	114
3	0.51	0.38	0.17	0.11	0.32	0.05	82,268	1,397	126
4	0.51	0.38	0.17	0.11	0.31	0.05	87,324	1,416	125
5	0.51	0.37	0.18	0.11	0.31	0.05	88,310	1,404	127
6	0.51	0.41	0.18	0.11	0.26	0.05	55,267	590	114
7	0.51	0.41	0.18	0.10	0.27	0.05	60,578	495	111
8	0.50	0.40	0.19	0.11	0.30	0.04	48,255	488	114
Sample of Districts that Opened Remotely in Fall 2020									
1	0.51	0.28	0.21	0.11	0.36	0.06	32,379	695	52
2	0.51	0.32	0.20	0.11	0.35	0.06	49,740	831	62
3	0.51	0.31	0.19	0.11	0.37	0.06	61,177	1,023	73
4	0.51	0.31	0.19	0.11	0.36	0.05	63,783	1,027	71
5	0.51	0.31	0.20	0.11	0.36	0.06	65,631	1,030	74
6	0.51	0.35	0.21	0.11	0.29	0.06	40,425	440	66
7	0.50	0.36	0.21	0.10	0.29	0.06	45,806	375	65
8	0.50	0.33	0.21	0.11	0.34	0.05	36,347	377	70
Sample of Districts that Opened In-Person in Fall 2020									
1	0.51	0.62	0.09	0.11	0.16	0.02	12,332	277	48
2	0.51	0.59	0.12	0.11	0.16	0.02	18,806	329	52
3	0.51	0.58	0.13	0.10	0.16	0.02	21,091	374	53
4	0.51	0.58	0.12	0.11	0.18	0.02	23,541	389	54
5	0.50	0.56	0.12	0.10	0.19	0.03	22,679	374	53
6	0.51	0.56	0.11	0.11	0.19	0.03	14,842	150	48
7	0.51	0.57	0.10	0.10	0.21	0.03	14,772	120	46
8	0.51	0.58	0.10	0.11	0.19	0.02	11,908	111	44

Note: Grade refers to the grade each cohort of students was enrolled in during fall 2020.

Table 3. MAP Growth Test Marginal Reliability by Grade, Subject, and Term

	In-person testers			Remote testers		
Grade	F19	W20	F20	F19	W20	F20
			Math			
1	0.90	0.92	0.94	0.91	0.93	0.96
2	0.93	0.93	0.95	0.94	0.94	0.96
3	0.95	0.95	0.95	0.95	0.95	0.96
4	0.95	0.94	0.95	0.95	0.95	0.96
5	0.95	0.95	0.96	0.96	0.96	0.96
6	0.96	0.96	0.95	0.96	0.96	0.96
7	0.96	0.96	0.96	0.96	0.96	0.97
8	0.97	0.97	0.97	0.97	0.97	0.97
			Reading			
1	0.87	0.91	0.94	0.88	0.92	0.96
2	0.93	0.94	0.96	0.94	0.95	0.97
3	0.96	0.96	0.96	0.96	0.96	0.96
4	0.96	0.95	0.96	0.96	0.96	0.96
5	0.95	0.95	0.95	0.96	0.96	0.96
6	0.95	0.95	0.95	0.96	0.95	0.96
7	0.95	0.95	0.95	0.96	0.95	0.96
8	0.95	0.95	0.95	0.96	0.96	0.96

Note: Grade refers to the grade each cohort of students was enrolled in during fall 2020.
F19=Fall 2019 test scores. W20=Winter 2020 test scores. F20=Fall 2020 test scores.
For each cohort, the F19 and W20 results refer to the prior grade while F20 corresponds to the grade shown in each row.

Table 4. Test-Retest Reliability by Grade, Subject, and Fall 2020 Reopening Status

	In-person testers			Remote testers		
Grade	F19-W20	F19-F20	W20-F20	F19-W20	F19-F20	W20-F20
			Math			
1	0.77	0.67	0.72	0.77	0.42	0.44
2	0.84	0.71	0.73	0.85	0.53	0.55
3	0.86	0.78	0.80	0.87	0.69	0.71
4	0.87	0.83	0.84	0.89	0.79	0.80
5	0.89	0.86	0.88	0.91	0.85	0.85
6	0.91	0.87	0.88	0.92	0.85	0.86
7	0.90	0.88	0.90	0.92	0.87	0.88
8	0.92	0.90	0.91	0.93	0.89	0.89
			Reading			
1	0.66	0.59	0.67	0.66	0.40	0.45
2	0.82	0.68	0.72	0.82	0.54	0.58
3	0.86	0.79	0.82	0.86	0.73	0.76
4	0.86	0.82	0.83	0.87	0.79	0.80
5	0.86	0.83	0.84	0.88	0.82	0.83
6	0.86	0.83	0.84	0.88	0.83	0.83
7	0.87	0.84	0.85	0.88	0.83	0.84
8	0.87	0.85	0.85	0.88	0.83	0.84

Note: Grade refers to the grade each cohort of students was enrolled in during fall 2020. F19=Fall 2019 test scores. W20=Winter 2020 test scores. F20=Fall 2020 test scores.

Table 5. Results from Regression Model Predicting Fall 2020 Test Scores

	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	Math							
Intercept	165.23 (0.65)***	175.97 (0.49)***	189.76 (0.32)***	199.16 (0.26)***	210.75 (0.24)***	212.87 (0.25)***	221.44 (0.26)***	226.10 (0.31)***
Fall 2019 RIT Score	0.85 (0.01)***	0.80 (0.01)***	0.75 (0.01)***	0.86 (0.00)***	0.94 (0.00)***	0.83 (0.00)***	0.93 (0.01)***	0.90 (0.01)***
Tested Remotely in Fall 2020	6.83 (0.26)***	4.80 (0.18)***	2.73 (0.13)***	1.76 (0.11)***	1.27 (0.11)***	1.32 (0.12)***	1.17 (0.12)***	1.36 (0.15)***
Black	-1.23 (0.47)**	-1.82 (0.29)***	-1.98 (0.21)***	-1.72 (0.18)***	-1.03 (0.18)***	-0.47 (0.22)*	-0.90 (0.24)***	-1.18 (0.27)***
Hispanic	-1.16 (0.38)**	0.19 (0.27)	-1.32 (0.20)***	-0.63 (0.16)***	-0.38 (0.15)*	-0.91 (0.18)***	-1.19 (0.19)***	-1.19 (0.23)***
Asian	3.44 (1.03)***	3.18 (0.63)***	1.88 (0.45)***	1.81 (0.36)***	1.90 (0.33)***	1.51 (0.40)***	2.12 (0.42)***	1.17 (0.55)*
Other Race	-0.31 (0.44)	-0.80 (0.32)*	-0.67 (0.24)**	-0.79 (0.20)***	-0.53 (0.19)**	-0.94 (0.21)***	-0.66 (0.24)**	-0.55 (0.28)*
Fall 2019 RIT by Tested Remotely	-0.24 (0.02)***	-0.20 (0.01)***	-0.07 (0.01)***	-0.05 (0.01)***	-0.04 (0.01)***	-0.01 (0.01)**	0.01 (0.01)*	0.05 (0.01)***
Black by Tested Remotely	0.67 (0.52)	0.58 (0.34)	-0.19 (0.25)	-0.16 (0.21)	-1.02 (0.20)***	-1.12 (0.25)***	-0.38 (0.26)	0.46 (0.31)
Hispanic by Tested Remotely	0.58 (0.44)	-0.50 (0.30)	-0.65 (0.23)**	-1.12 (0.18)***	-1.55 (0.17)***	-1.29 (0.21)***	-0.66 (0.22)**	0.28 (0.26)
Asian by Tested Remotely	4.30 (1.08)***	4.23 (0.67)***	2.30 (0.49)***	1.51 (0.40)***	1.28 (0.36)***	2.35 (0.44)***	1.96 (0.45)***	2.49 (0.60)***
Other Race by Tested Remotely	0.69 (0.55)	1.52 (0.41)***	0.71 (0.30)*	0.45 (0.25)	-0.30 (0.24)	0.35 (0.27)	0.13 (0.29)	-0.37 (0.36)
District Average SES	-0.11 (0.14)	0.58 (0.10)***	0.81 (0.07)***	0.65 (0.06)***	0.28 (0.06)***	0.89 (0.06)***	0.25 (0.06)***	0.31 (0.07)***
% SPED in District	-25.77 (4.24)***	-8.04 (3.25)*	-10.85 (2.13)***	-3.05 (1.78)	-11.43 (1.70)***	8.75 (1.71)***	2.01 (1.74)	-0.65 (2.03)
% ELL in District	-3.83 (1.18)**	-2.83 (0.89)**	-5.94 (0.44)***	-4.15 (0.36)***	-3.91 (0.34)***	-1.26 (0.37)***	-1.61 (0.37)***	-2.15 (0.40)***
Suburb	-0.20 (0.17)	-1.49 (0.11)***	-1.32 (0.08)***	-0.52 (0.07)***	-1.13 (0.06)***	-0.35 (0.09)***	-1.46 (0.09)***	-2.28 (0.11)***
Town	-3.16 (0.29)***	-2.33 (0.23)***	-2.53 (0.17)***	-1.51 (0.14)***	-1.60 (0.14)***	-0.30 (0.14)*	-1.37 (0.14)***	-1.34 (0.16)***
Rural	-0.18 (0.28)	-1.39 (0.21)***	-1.88 (0.15)***	-0.94 (0.13)***	-0.91 (0.12)***	-0.19 (0.14)	-0.58 (0.14)***	-1.39 (0.16)***
	Reading							
Intercept	158.20 (0.64)***	173.68 (0.58)***	189.04 (0.39)***	198.99 (0.30)***	206.21 (0.28)***	212.01 (0.30)***	216.24 (0.30)***	220.38 (0.30)***
Fall 2019 RIT Score	0.86 (0.02)***	0.93 (0.01)***	0.82 (0.01)***	0.80 (0.00)***	0.81 (0.00)***	0.80 (0.01)***	0.83 (0.01)***	0.83 (0.01)***
Tested Remotely in Fall 2020	6.05 (0.26)***	4.97 (0.21)***	2.20 (0.16)***	0.83 (0.13)***	0.45 (0.12)***	-0.23 (0.14)	-0.14 (0.14)	-0.37 (0.15)*
Black	-1.77 (0.46)***	-1.04 (0.35)**	-1.69 (0.25)***	-2.25 (0.21)***	-1.61 (0.20)***	-1.34 (0.24)***	-1.23 (0.28)***	-1.59 (0.28)***
Hispanic	-2.56 (0.40)***	-0.84 (0.32)**	-1.69 (0.24)***	-1.13 (0.18)***	-0.83 (0.17)***	-1.25 (0.21)***	-1.59 (0.22)***	-1.17 (0.22)***
Asian	2.45 (1.08)*	3.04 (0.75)***	0.18 (0.55)	0.40 (0.42)	0.81 (0.37)*	1.77 (0.47)***	0.94 (0.47)*	1.00 (0.47)*
Other Race	-0.94 (0.43)*	-0.86 (0.38)*	-0.77 (0.29)**	-0.64 (0.24)**	-0.93 (0.22)***	-1.12 (0.24)***	-0.36 (0.27)	-0.22 (0.28)
Fall 2019 RIT by Tested Remotely	-0.21 (0.02)***	-0.21 (0.01)***	-0.08 (0.01)***	-0.03 (0.01)***	0.00 (0.00)	0.01 (0.01)*	0.02 (0.01)**	0.04 (0.01)***
Black by Tested Remotely	0.57 (0.51)	0.11 (0.40)	-0.69 (0.29)*	-0.53 (0.24)*	-0.74 (0.23)**	-0.83 (0.28)**	-0.88 (0.31)**	0.27 (0.31)
Hispanic by Tested Remotely	-0.05 (0.45)	-0.54 (0.37)	-1.43 (0.27)***	-2.03 (0.21)***	-1.92 (0.20)***	-1.27 (0.25)***	-0.96 (0.25)***	0.01 (0.25)
Asian by Tested Remotely	3.55 (1.14)**	3.58 (0.81)***	1.27 (0.59)*	0.30 (0.47)	0.34 (0.41)	-0.07 (0.52)	0.75 (0.50)	1.57 (0.51)**
Other Race by Tested Remotely	1.67 (0.53)**	2.17 (0.47)***	0.66 (0.35)	-0.28 (0.30)	0.23 (0.27)	0.51 (0.32)	-0.78 (0.33)*	-0.36 (0.35)
District Average SES	0.81 (0.12)***	0.81 (0.10)***	1.26 (0.07)***	0.86 (0.07)***	0.80 (0.06)***	0.62 (0.07)***	0.40 (0.06)***	0.49 (0.07)***
% SPED in District	-13.28 (4.08)**	-18.51 (3.76)***	-4.78 (2.53)	-3.62 (2.04)	-1.48 (1.94)	4.78 (2.00)*	5.15 (1.94)**	-0.41 (1.89)
% ELL in District	-0.39 (1.09)	0.95 (1.01)	-3.44 (0.52)***	-2.93 (0.42)***	-2.37 (0.38)***	-1.33 (0.42)**	0.61 (0.42)	-0.95 (0.42)*
Suburb	2.59 (0.17)***	1.43 (0.14)***	-0.51 (0.10)***	-0.06 (0.08)	-0.58 (0.07)***	-0.24 (0.10)*	-0.10 (0.09)	-0.52 (0.10)***
Town	-1.94 (0.28)***	-1.29 (0.27)***	-2.28 (0.19)***	-1.32 (0.17)***	-0.62 (0.16)***	-0.42 (0.16)*	-0.35 (0.17)*	-0.28 (0.17)
Rural	1.73 (0.28)***	0.20 (0.24)	-1.00 (0.18)***	-0.50 (0.15)**	-0.50 (0.14)***	0.09 (0.16)	0.33 (0.16)*	-0.07 (0.16)

Note: Grade refers to the grade each cohort of students was enrolled in during fall 2020.

Figure 1: Trends in Average Response Time Effort (RTE) in Reading by Grade and Fall 2020 Reopening Status

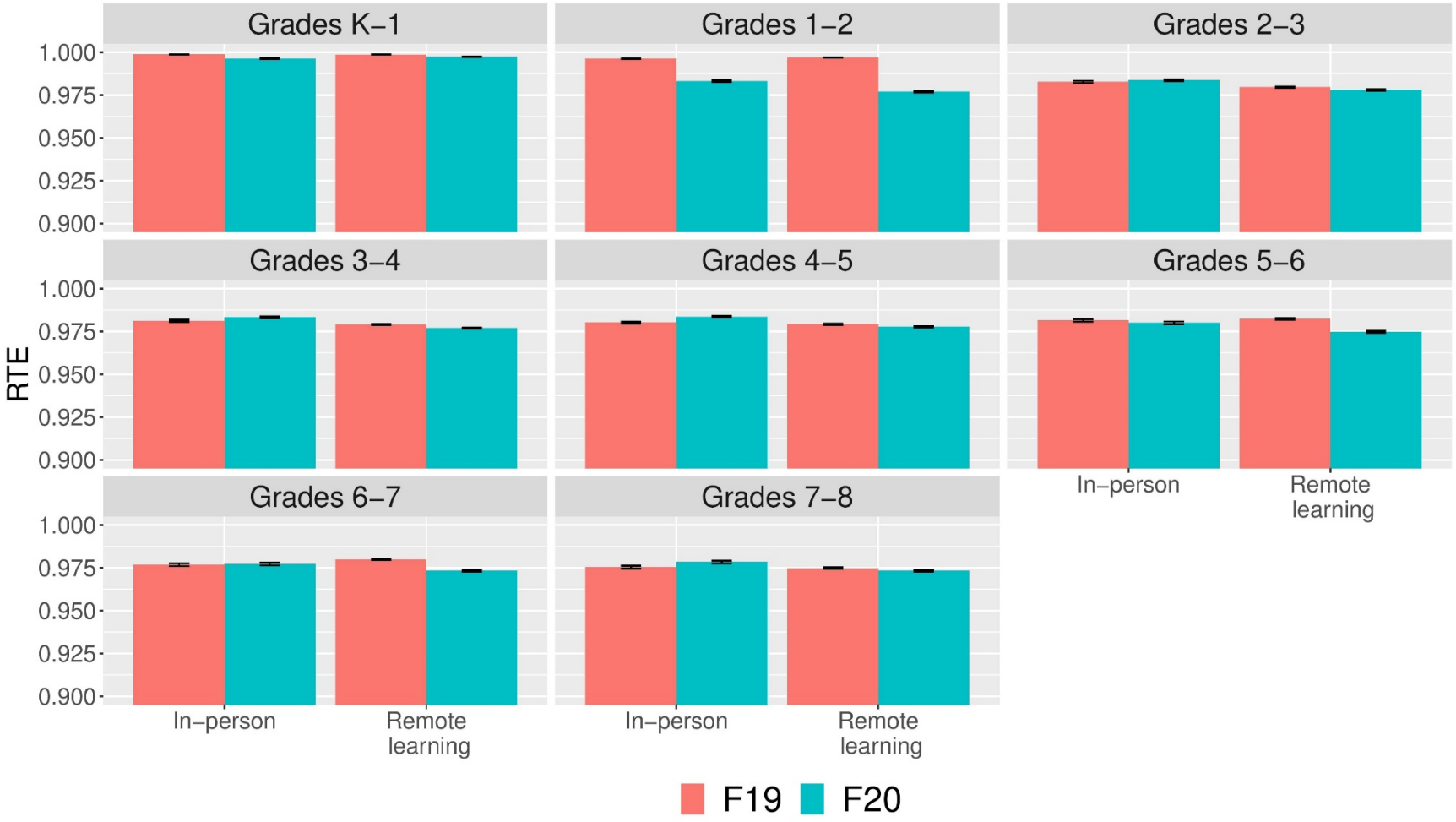


Figure 2: Trends in Average Test Duration in Reading by Grade and Fall 2020 Reopening Status

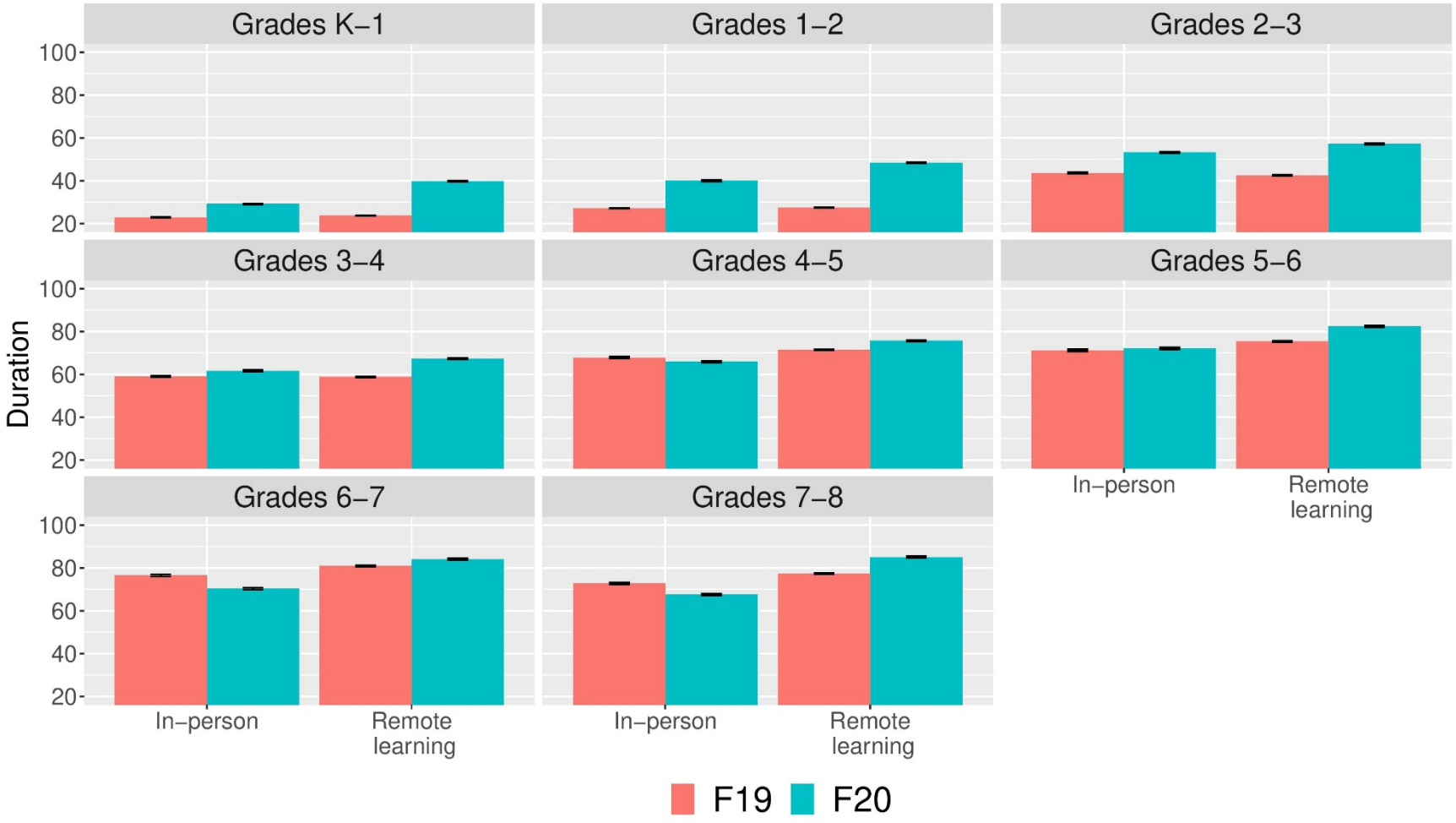


Figure 3: Average Changes in Test Score Percentiles Between Fall 2019 and Fall 2020 in Math by Grade and Fall 2020 Reopening Status

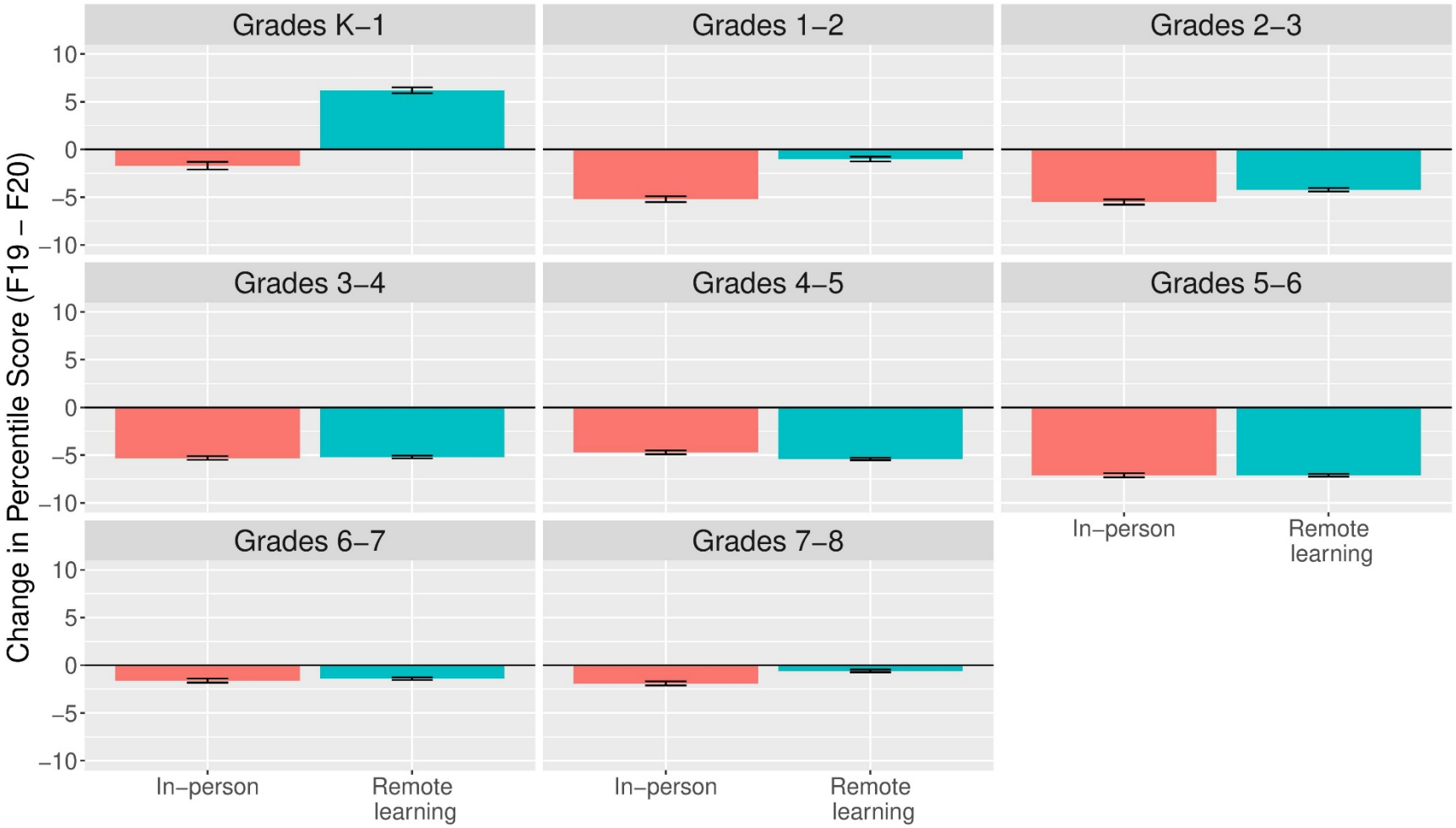


Figure 4: Average Changes in Test Score Percentiles Between Fall 2019 and Fall 2020 in Reading by Grade and Fall 2020 Reopening Status

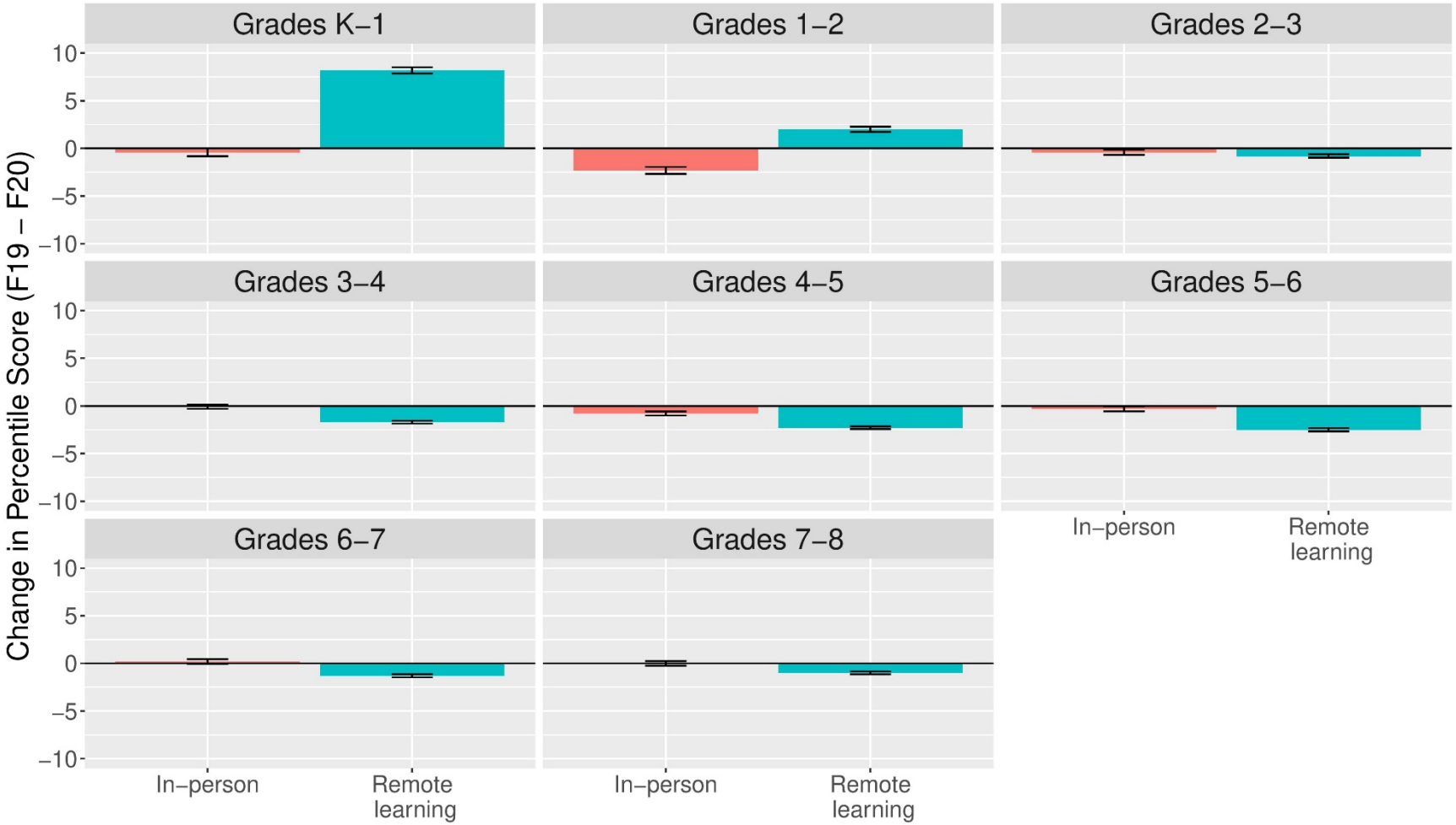
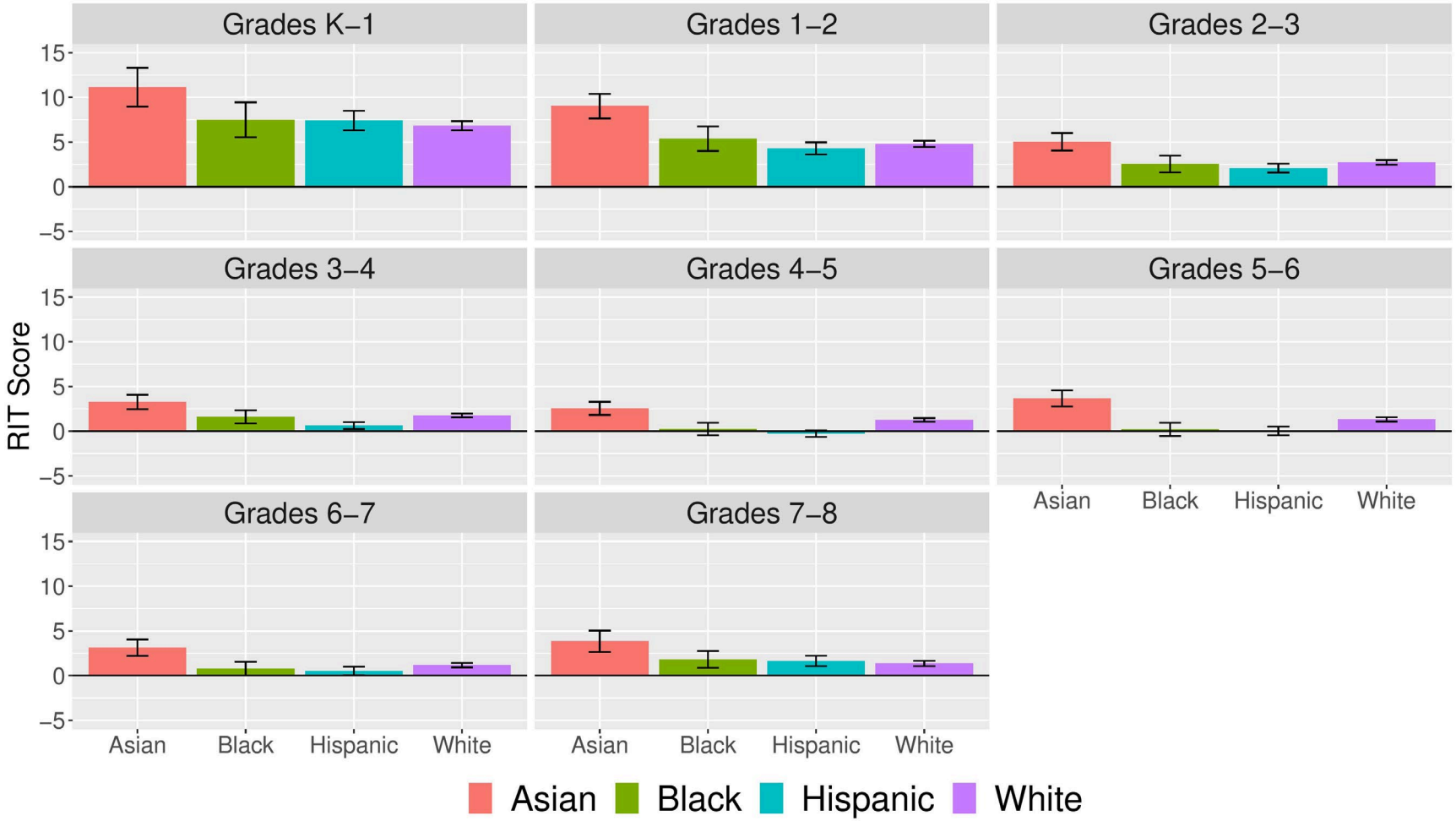
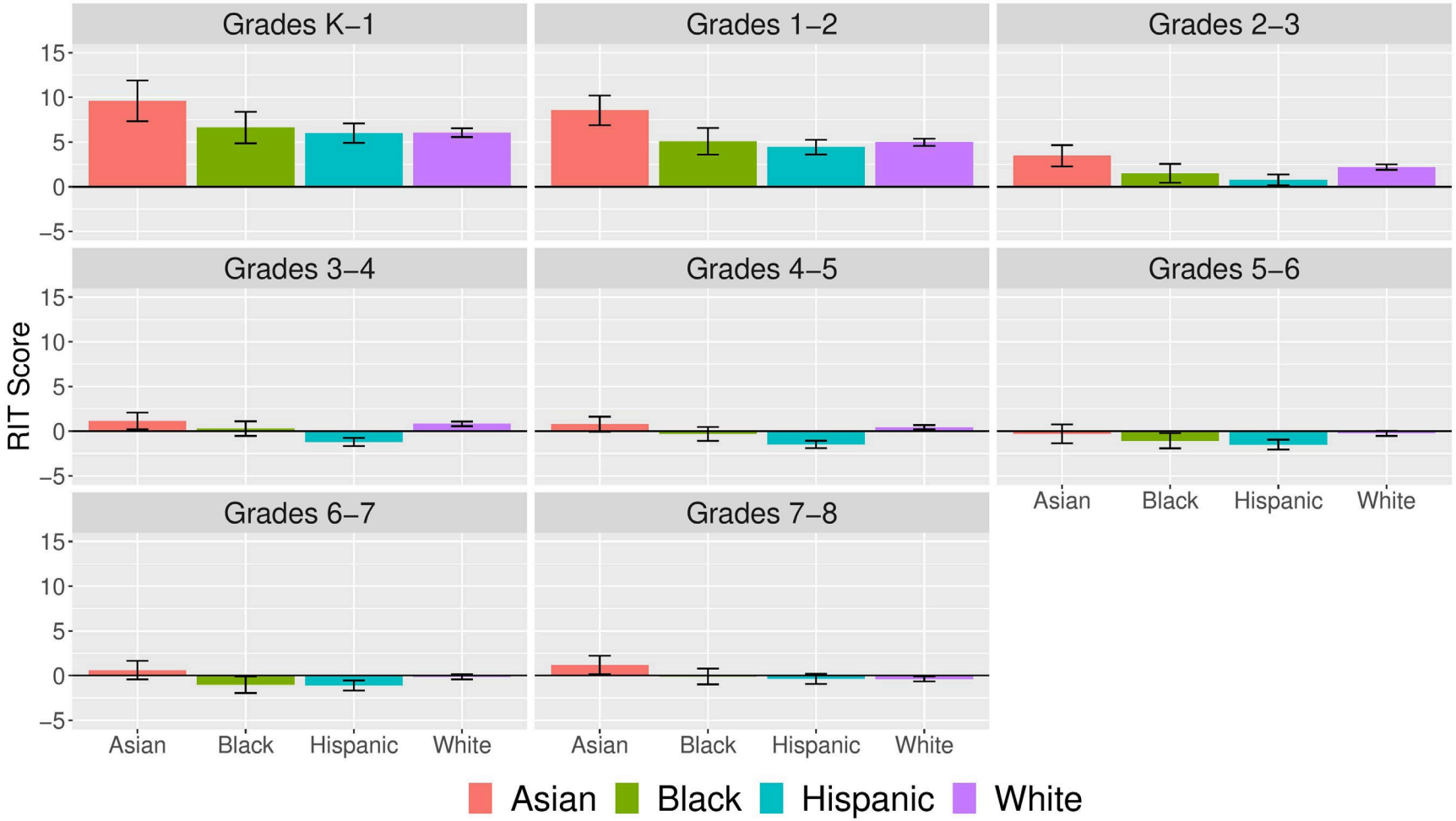


Figure 5: Average Difference in Fall 2020 Math RIT Scores Between Remote and In-Person Testers by Grade and Racial/Ethnic Groups (Controlling for Prior Achievement and District Characteristics)



Note: Positive values indicate remote testers scored better on average than their same-race peers within in-person settings, while negative values indicate in-person testers scored better on average in fall 2020. Reported estimates were calculated based on the Remote variable and the Remote by race/ethnicity interaction terms.

Figure 6: Average Difference in Fall 2020 Reading RIT Scores Between Remote and In-Person Testers by Grade and Racial/Ethnic Groups (Controlling for Prior Achievement and District Characteristics)



Note: Positive values indicate remote testers scored better on average than their same-race peers in in-person settings, while negative values indicate in-person testers scored better on average in fall 2020. Reported estimates were calculated based on the Remote variable and the remote by race/ethnicity interaction terms.

7. References

[i] Meyer, P. (2020). Comparisons between remote testing and in-school testing for MAP Growth: A summary of results for spring 2020. NWEA.

[ii] Cronin, J. & Wise, S. (2020). Assessment in the time of COVID-19: Engagement during remote low-stakes testing. Unpublished manuscript. NWEA.

[iii] Johnson, A. & Kuhfeld, M. (2020). Fall 2019 to fall 2020 MAP Growth attrition analysis. NWEA. https://www.nwea.org/research/publication/fall-2019-to-fall-2020-map-growth-attrition-analysis/

[iv] School Districts' Reopening Plans: A Snapshot (2020, July 15). *Education Week*. Retrieved September 28, 2020 from https://www.edweek.org/ew/section/multimedia/school-districts-reopening-plans-a-snapshot.html

[v] Thum, Y. M., & Kuhfeld, M. (2020). NWEA 2020 MAP Growth achievement status and growth norms for students and schools. NWEA Research Report. Portland, OR: NWEA.

[vi] Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*, 237-252.

[vii] Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52-61.

[viii] NWEA. (2019). MAP® Growth™ Technical Report. Portland, OR: NWEA.

[ix] Kuhfeld, M., Tarasawa, B., Johnson, A., Ruzek, E., & Lewis, K. (2020). Learning during COVID-19: Initial findings on students' reading and math achievement and growth. NWEA. https://www.nwea.org/research/publication/learning-during-covid-19-initial-findings-on-students-reading-and-math-achievement-and-growth/