EOG and EOC Mathematics Scores in 2020-21 after COVID-19 Related Disruptions to Full-Time, In-Person Learning

Introduction

The purpose of this study was to investigate the impact of interruptions to full-time, in-person instruction due to COVID-19 on student learning, particularly performance on end-of-grade (EOG) and end-of-course (EOC) mathematics tests. The data at hand consisted of test results for EOG mathematics Grades 4 and 7 and EOC NC Math 1 for the 2019 and 2021 school years. The data were further subsetted by region, sex, and ethnicity in order to assess the differential effect, if any, among various subgroups across the state. Table 1 below contains a summary of the sample sizes overall and for each subgroup. In general, sample sizes for the two EOGs and NC Math 1 were smaller for the overall population and each subgroup for 2021 compared to 2019. Tables 2 to 4 and Figure 1 contain data on historic trends in sample sizes from the 2014-15 administration to present. The year-over-year changes appear to be random with no clear trend. Therefore, there is no evidence to suggest that the drop in sample size for 2021 was part of a broader historical trend.

Table 1. Sample Sizes for the Overall Population and by Subgroup

		Grade 4		Grade 7		NC Math 1	
		2019	2021	2019	2021	2019	2021
Overall		120,416	105,979	118,483	111,894	99,725	103,238
Region	North Central	30,392	25,420	29,751	25,586	23,444	25,166
	Northeast	5,685	4,842	5,680	5,446	5,185	4,858
	Northwest	6,222	5,476	6,418	6,304	5,024	5,992
	Piedmont-Triad	19,561	17,360	19,180	18,416	12,826	13,712
	Sandhills	10,783	9,659	10,388	10,123	9,982	9,378
	Southeast	11,134	10,501	10,843	10,810	9,489	10,349
	Southwest	29,756	26,341	29,275	28,115	27,800	27,454
	Western	6,601	6,071	6,463	6,607	5,975	5,902
Sex	Female	58,536	51,615	57,887	54,565	48,772	50,892
	Male	61,782	53,780	60,587	56,999	50,320	51,954
Ethnicity	White	55,056	47,633	55,748	50,974	49,776	49,705
	Black	30,773	25,624	29,574	27,591	23,556	24,111
	Hispanic/Latino	23,125	21,014	22,319	22,370	16,929	19,354
	Other	11,364	11,124	10,833	10,629	8,831	9,675

Table 2. Historic Sample Sizes for Grade 4

		2015	2016	2017	2018	2019	2021
Overall		113,968	116,792	120,487	121,545	120,416	105,979
Sex	Female	55,849	57,162	58,728	59,471	58,536	51,615
	Male	58,119	59,630	61,759	62,074	61,782	53,780
Ethnicity	White	56,960	56,473	56,921	56,423	55,056	47,633
	Black	28,677	29,664	30,382	31,090	30,773	25,624
	Hispanic/Latino	19,151	20,629	22,391	22,838	23,125	21,014
	Other	9,180	10,026	10,793	11,194	11,364	11,124

Table 3. Historic Sample Sizes for Grade 7

		2015	2016	2017	2018	2019	2021
Overall		114,662	115,005	107,174	115,288	118,483	111,894
Sex	Female	55,933	56,076	52,059	56,427	57,887	54,565
	Male	58,729	58,929	55,115	58,861	60,587	56,999
Ethnicity	White	58,514	58,297	53,853	56,115	55,748	50,974
	Black	29,622	28,791	26,400	28,620	29,574	27,591
	Hispanic/Latino	17,371	18,234	17,616	20,611	22,319	22,370
	Other	9,155	9,683	9,305	9,942	10,833	10,629

Table 4. Historic Sample Sizes for NC Math 1

		2015	2016	2017	2018	2019	2021
Overall		118,802	125,087	122,141	119,674	99,725	103,238
Sex	Female	57,519	60,532	59,099	58,371	48,772	50,892
	Male	61,283	64,555	63,042	61,303	50,320	51,954
Ethnicity	White	60,870	62,405	60,790	58,806	49,776	49,705
	Black	32,125	33,510	31,160	30,058	23,556	24,111
	Hispanic/Latino	16,452	19,105	19,983	20,254	16,929	19,354
	Other	9,355	10,067	10,208	10,556	8,831	9,675

Figure 1. Historic Sample Sizes for the Overall Population
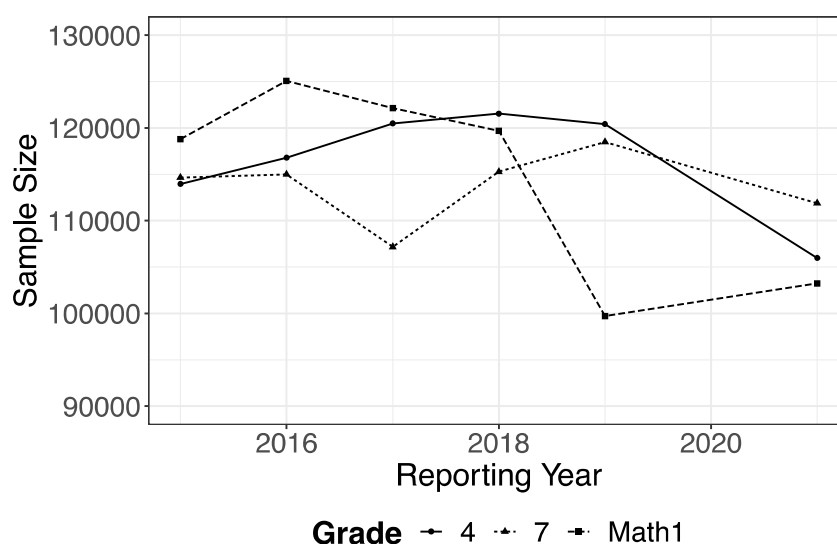
Descriptive Statistics

To understand the impact COVID-19 had on 2020-21 school year in North Carolina, student EOG and EOC test data from Public School Units (PSU) were compared between 2019 and 2021 using two variables: scale scores and achievement levels. Results for the scale scores are presented first followed by a discussion of results for the achievement levels.

Scale Scores

Overall Population. Student scale scores from all PSU were analyzed to compare the difference between the average student scores for 2019 and 2021. On average, scale scores for the 2021 student population were about half standard deviation lower than those for the 2019 student population (see

Tables 5 to 7). Figure 2 contains histograms showing the distributions of the relative frequencies of scale scores by year for the two EOGs and NC Math 1. The dashed lines represent the mean scale score for each year. For all three assessments, the scale scores from 2021were shifted to the left of the scale scores from 2019, indicating a less proficient examinee population. Furthermore, the mean differences in scale scores across the two years were statistically significant for the two EOGs and NC Math 1. Generally, the term "statistically significant" means that a difference, no matter how small, is not likely due to chance. Therefore, statistical significance in the current study implies that the 2019 and 2021 student populations were statistically distinct in terms of their mean scale scores due to some other factors than random chance, such as learning interruptions caused by COVID-19.

Although statistical tests are commonly used in practice to compare the means of two or more populations, sampling error generally decreases with the increase of sample size, making it more and more likely to find significant results. Accordingly, effect size calculations are often used to quantify the practical significance of a mean difference. The effect size measure called Cohen's d was calculated by dividing the mean difference of the two samples by their pooled standard deviation (which is a measure of the variability or spread of the underlying scores). This allows for expressing the mean difference of the scores in units of standard deviation. A common rule of thumb used in interpreting Cohen's d is that $d = .2$ be considered a 'small' effect size, .5 be considered a 'medium' effect size, and .8 be considered a 'large' effect size; however, in the educational context, .2 is often considered a nontrivial effect size.

As can be seen from Tables 5 to 7, all effect sizes were greater than .2. The largest overall effect size was observed for Grade 4, which was .5, indicating that the mean difference in scale scores between the student populations from 2019 and 2021 was about a half standard deviation. Compared to Grade 4, smaller effect sizes were observed for Grade 7 and NC Math 1; however, the effect sizes were still meaningful with values close to .4 (see Tables 6 and 7). Table 8 contains historic scale score

means, standard deviations, and year-over-year effective sizes from 2015 to 2018 (the scale changed in the 2019 administration, so the effect size from 2018 to 2019 is not meaningful). Historically, effect sizes were close to zero. In this context, all of the effect size measures observed between 2019 and 2021 were larger than historical year-over-year measures.

Another way of analyzing overall scale score differences is to compare the scale score needed for a student to be at a given percentile in 2019 versus 2021. A percentile is the scale score at which or below a given percentage of examinees falls. For example, 50 percent of students have scale scores at or below the 50th percentile scale score. Table 9 contains the 25th, 50th, and 75th percentile for 2019 and 2021, as well as the difference. Taking the 50th percentile as an example, for Grade 4, students had to earn a scale score of 549 to be at the 50th percentile in 2019, whereas based on the 2021 cohort, the 50th percentile corresponded to a scale score of 542, a 7-point difference (.7 standard deviation). Similarly, for Grade 7 and NC Math 1, a 5 point lower scale score was required in 2021 to be at the 50th percentile than 2019. Figure 3 contains boxplots of the scale scores for 2019 and 2021 by grade. The line inside each box represents the 50th percentile, while the edge of each box represents the 25th percentile and the 75th percentile, respectively. The boxplots are shifted downward in 2021, which is consistent with the rest of the analysis of overall scale scores.

Subgroups. Consistent with the results observed for the overall population, scale scores for the 2021 student populations were, on average, lower than those for the 2019 student populations for all subgroups (Appendix A in a separate file contains the histograms comparing each relevant subgroup by testing year for each grade level). Moreover, the mean differences in scale scores between the 2019 and 2021 student populations were statistically significant. As shown in Tables 5 to 7, effect sizes for all subgroups were above .2 across the two EOGs and NC Math 1, indicating that the mean difference was not only statistically significant but also practically significant in the educational context.

In terms of region, Sandhills had the largest effect size (.7) for Grade 4, followed by Piedmont-

Triad (.5), North Central (.5), and Northeast (.5). For Grade 7, Western (.5) and Sandhills (.5) showed larger effect sizes than the other regions. Sandhills (.5) again showed the largest effect size for NC Math 1, along with Piedmont-Triad (.5). For the two EOGs and NC Math 1, both male and female students showed effect sizes larger than .2, with female students showing somewhat larger effect sizes than male students. Finally, all ethnic groups showed effect sizes larger than .2. In general, Black and Hispanic/Latino students tended to show larger effect sizes than white students for both EOGs and NC Math 1.

Table 5. Grade 4: Mean Differences in Scale Scores and Effect Sizes by Subgroup

		2019		2021		Mean	Effect
		Mean	SD	Mean	SD	Diff.	Size
Overall		548.3	10.0	543.4	9.8	4.9	0.5
Region	North Central	548.7	10.2	543.5	10.0	5.1	0.5
	Northeast	546.9	9.7	542.1	9.5	4.8	0.5
	Northwest	548.5	9.6	544.0	9.4	4.5	0.5
	Piedmont-Triad	547.8	10.1	542.7	9.7	5.0	0.5
	Sandhills	546.3	9.5	540.2	8.5	6.1	0.7
	Southeast	547.4	9.8	543.4	9.6	4.0	0.4
	Southwest	549.4	10.1	544.9	10.2	4.5	0.4
	Western	549.1	9.8	544.6	9.6	4.5	0.5
Sex	Female	548.0	9.8	542.9	9.6	5.2	0.5
	Male	548.6	10.3	544.0	10.1	4.5	0.4
Ethnicity	White	551.3	9.5	546.9	9.7	4.3	0.5
	Black	543.6	8.9	538.2	7.4	5.4	0.6
	Hispanic/Latino	546.4	9.4	540.8	8.5	5.6	0.6
	Other	550.6	10.7	545.8	10.8	4.7	0.4

Table 6. Grade 7: Mean Differences in Scale Scores and Effect Sizes by Subgroup

		2019		2021		Mean	Effect
		Mean	SD	Mean	SD	Diff.	Size
Overall		548.1	9.7	544.3	8.9	3.8	0.4
Region	North Central	548.4	9.9	544.7	9.2	3.8	0.4
	Northeast	546.8	9.4	543.2	8.5	3.6	0.4
	Northwest	548.6	9.2	545.0	8.5	3.6	0.4
	Piedmont-Triad	547.5	9.6	543.5	8.5	4.0	0.4
	Sandhills	545.8	9.0	541.8	7.5	4.0	0.5
	Southeast	546.8	9.2	543.9	8.4	2.7	0.3
	Southwest	549.5	10.1	545.8	9.5	3.8	0.4
	Western	548.8	9.2	544.5	8.3	4.3	0.5
Sex	Female	548.5	9.6	544.3	8.8	4.3	0.5
	Male	547.7	9.8	544.5	9.0	3.3	0.3
Ethnicity	White	551.0	9.4	547.1	8.9	3.9	0.4
	Black	543.4	8.2	540.1	6.7	3.4	0.4
	Hispanic/Latino	545.8	8.8	542.2	7.5	3.6	0.4
	Other	550.8	10.8	547.0	10.6	3.8	0.4

Table 7. NC Math 1: Mean Differences in Scale Scores and Effect Sizes by Subgroup

		2019		2021		Mean	Effect
		Mean	SD	Mean	SD	Diff.	Size
Overall		549.8	9.5	546.2	9.1	3.7	0.4
Region	North Central	549.8	9.8	545.7	9.0	4.1	0.4
	Northeast	547.5	9.1	544.8	8.7	2.7	0.3
	Northwest	549.8	8.6	548.1	8.8	1.7	0.2
	Piedmont-Triad	549.8	9.2	545.7	8.4	4.1	0.5
	Sandhills	547.7	9.0	543.4	8.0	4.3	0.5
	Southeast	549.3	9.1	545.3	8.5	4.0	0.5
	Southwest	551.0	10.0	547.5	9.7	3.5	0.4
	Western	550.8	8.7	547.8	8.5	3.0	0.4
Sex	Female	550.5	9.3	546.3	9.0	4.1	0.5
	Male	549.2	9.7	546.0	9.1	3.2	0.3
Ethnicity	White	552.2	9.1	548.6	8.9	3.6	0.4
	Black	545.3	8.4	541.7	7.1	3.6	0.5
	Hispanic/Latino	547.7	8.9	544.1	8.0	3.6	0.4
	Other	552.5	10.7	548.8	10.7	3.7	0.3

Table 8. 2015-2018 Scale Score Means, Standard Deviations, and Effect Sizes

	2015	2016	2017	2018
Grade 4				
Mean	449.7	449.9	450.0	449.9
SD	10.0	10.0	10.1	10.1
Effect Size		0.0	0.0	0.0
Grade 7				
Mean	449.7	450.2	450.5	450.8
SD	10.1	10.3	10.3	10.3
Effect Size		0.0	0.0	0.0
Math 1				
Mean	250.2	250.6	251.9	251.6
SD	10.0	10.3	9.9	10.0
Effect Size		0.0	-0.1	0.0

Table 9. 2019 and 2021 Scale Score Percentiles

	25%	50%	75%
Grade 4			
2019	541	549	556
2021	535	542	551
Difference	6	7	5
Grade 7			
2019	540	548	555
2021	537	543	550
Difference	3	5	5
Math1			
2019	542	550	556
2021	539	545	552
Difference	3	5	4

Figure 2. Histograms of Scale Scores for the Overall Population
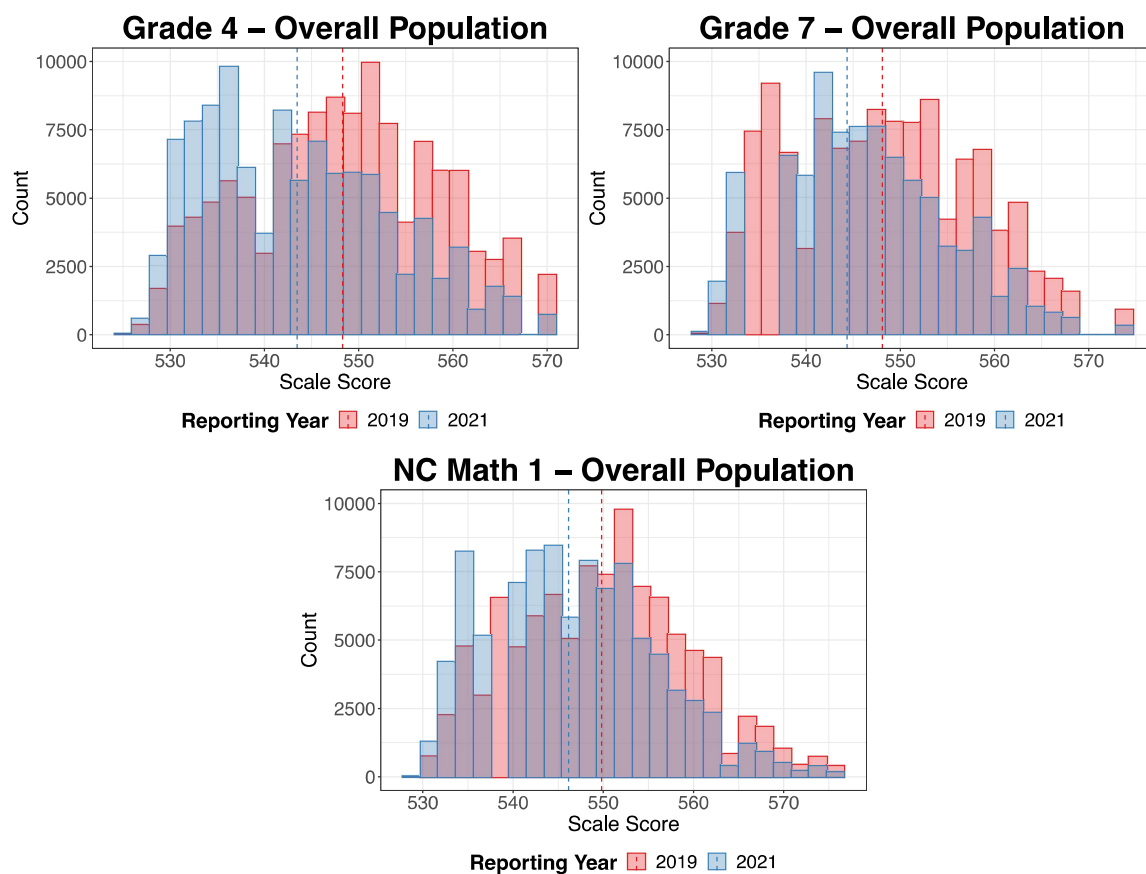
Figure 3. Boxplots of Scale Scores for the Overall Population.
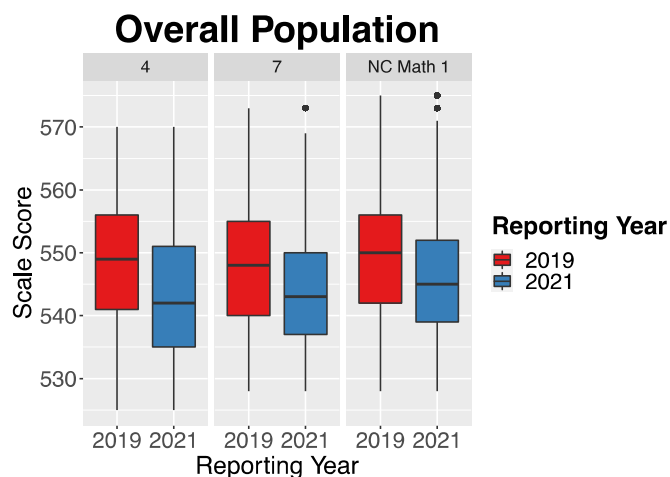
Achievement Levels

Overall Population. Tables 10 to 12 contain the percentages of students classified as grade-level

proficient (i.e., Levels 3, 4, and 5) across the two years for the two EOGs and NC Math 1. Additionally,

Figures 4 to 7 contain bar graphs showing the relative percentages of students classified as grade-level proficient for the overall population and each subgroup. Overall, the percentages of students classified as grade-level proficient for the 2021 populations were lower than those for the 2019 populations. Similar to scale scores, the differences in percentages between 2019 and 2021 were statistically significant. From Table 10, it can be seen that the percentage difference for Grade 4 overall was 21%, a meaningful difference. Compared to Grade 4, smaller percentage differences were observed for Grade 7 and NC Math 1; however, the differences were still large with values close to 17% (see Tables 11 and 12).

Subgroups. Consistent with the overall results, the percentages of students classified as grade-level proficient for the 2021 student populations were lower than those for the 2019 student populations, and the differences were statistically significant. In general, larger drops were observed for Grade 4 compared to Grade 7 and NC Math 1. For all three assessments, Sandhills showed the largest percentage differences, and female students showed larger percentage differences than male students. In terms of ethnicity, Black and Hispanic/Latino students showed larger percentage drops than White students.

Table 10. Grade 4: Percentage of Students Classified as Grade-Level Proficient for the Overall Population and by Subgroups

		2019	2021	Diff.
Overall		57.4%	36.4%	21.0%
Region	North Central	58.5%	36.8%	21.7%
	Northeast	51.2%	30.8%	20.4%
	Northwest	59.5%	39.1%	20.4%
	Piedmont-Triad	55.1%	33.4%	21.7%
	Sandhills	49.6%	22.9%	26.7%
	Southeast	54.2%	36.2%	18.0%
	Southwest	62.1%	42.7%	19.4%
	Western	61.8%	41.6%	20.2%
Sex	Female	56.6%	34.2%	22.4%
	Male	58.3%	38.8%	19.5%
Ethnicity	White	70.3%	51.6%	18.7%
	Black	37.0%	14.3%	22.7%
	Hispanic/Latino	50.8%	24.8%	26.0%
	Other	64.5%	45.3%	19.2%

Table 11. Grade 7: Percentage of Students Classified as Grade-Level Proficient for the Overall Population and by Subgroups

		2019	2021	Diff.
Overall		58.5%	41.4%	17.1%
Region	North Central	59.2%	42.3%	16.9%
	Northeast	53.5%	35.2%	18.3%
	Northwest	62.4%	45.7%	16.7%
	Piedmont-Triad	56.1%	37.7%	18.4%
	Sandhills	48.8%	29.1%	19.7%
	Southeast	53.8%	39.6%	14.2%
	Southwest	63.7%	47.5%	16.2%
	Western	64.0%	44.4%	19.6%
Sex	Female	60.6%	40.8%	19.8%
	Male	56.5%	41.9%	14.6%
Ethnicity	White	71.5%	55.7%	15.8%
	Black	37.8%	20.0%	17.8%
	Hispanic/Latino	49.8%	31.1%	18.7%
	Other	65.8%	50.2%	15.6%

Table 12. NC Math 1: Percentage of Students Classified as Grade-Level Proficient for the Overall Population and by Subgroups

		2019	2021	Diff.
Overall		60.2%	42.8%	17.4%
Region	North Central	59.1%	40.4%	18.7%
	Northeast	49.8%	36.6%	13.2%
	Northwest	62.3%	52.7%	9.6%
	Piedmont-Triad	60.9%	40.8%	20.1%
	Sandhills	51.4%	30.2%	21.2%
	Southeast	58.4%	39.7%	18.7%
	Southwest	64.5%	48.4%	16.1%
	Western	67.0%	52.1%	14.9%
Sex	Female	63.4%	43.6%	19.8%
	Male	57.1%	42.1%	15.0%
Ethnicity	White	71.3%	55.4%	15.9%
	Black	39.9%	21.2%	18.7%
	Hispanic/Latino	52.0%	33.0%	19.9%
	Other	67.1%	51.9%	15.2%

Figure 4. Overall Population: Percentage of Students Classified as Grade-Level Proficient for the Two EOGs and NC Math 1

Overall Population
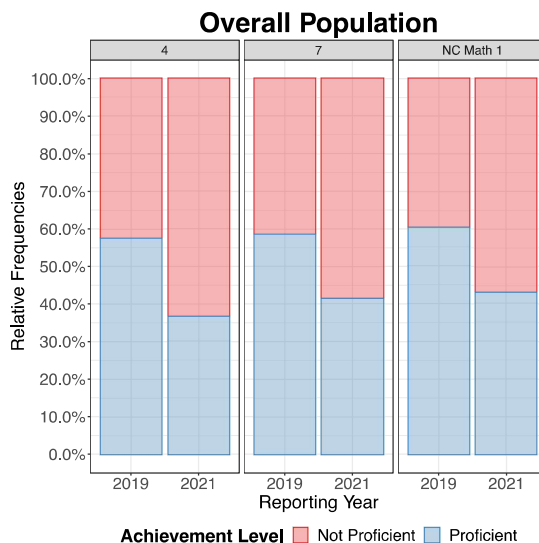
Figure 5. Region: Percentage of Students Classified as Grade-Level Proficient for the Two EOGs and NC Math 1
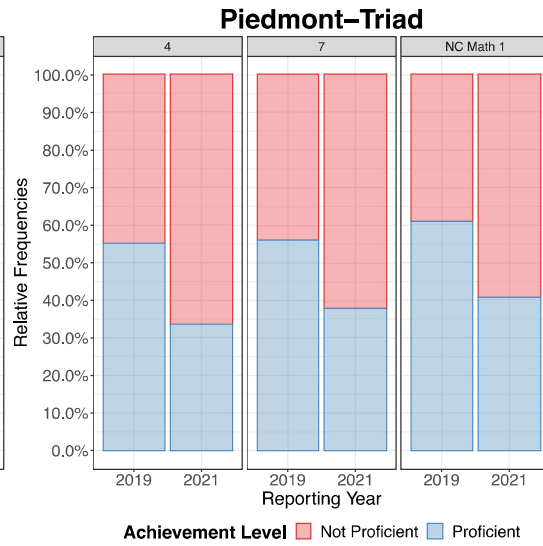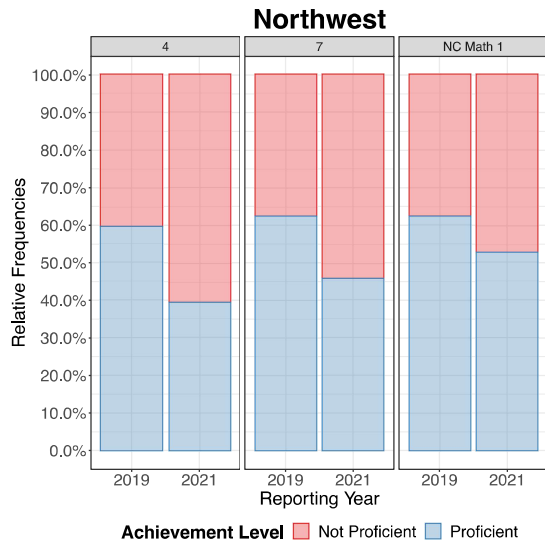
North Central

Northeast
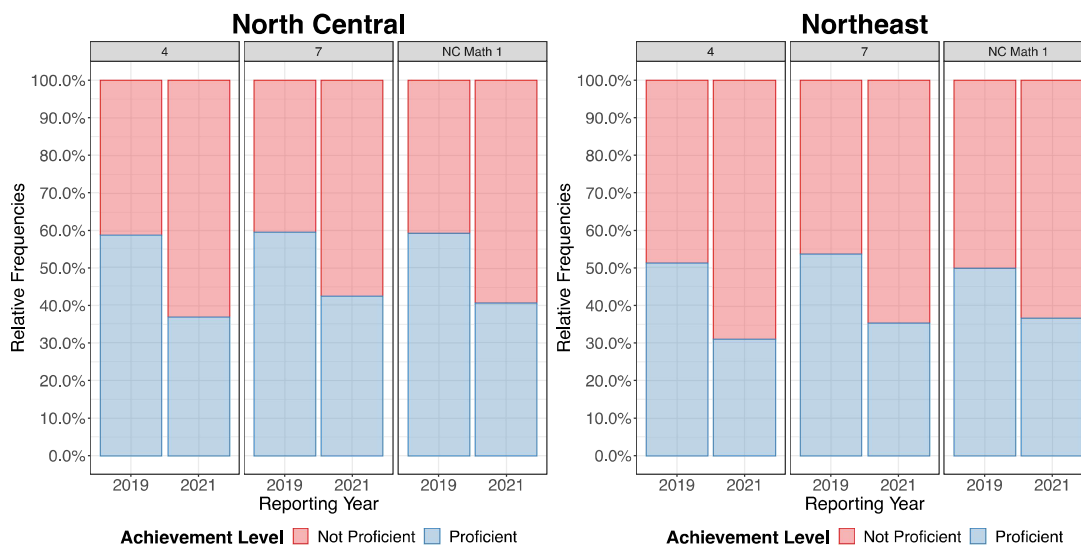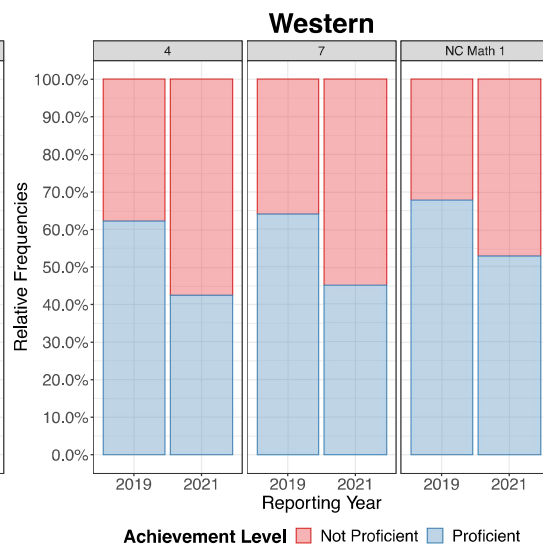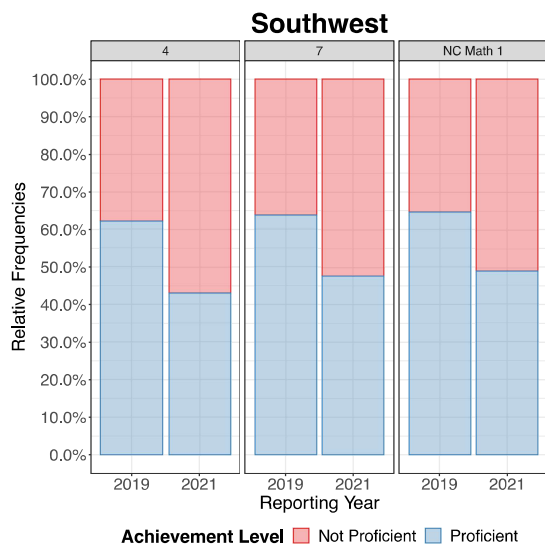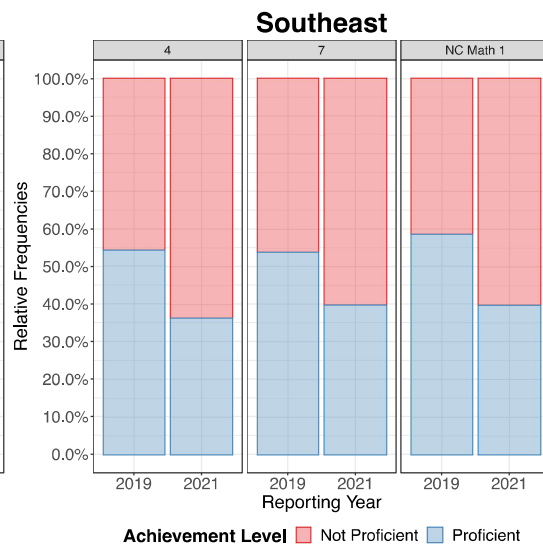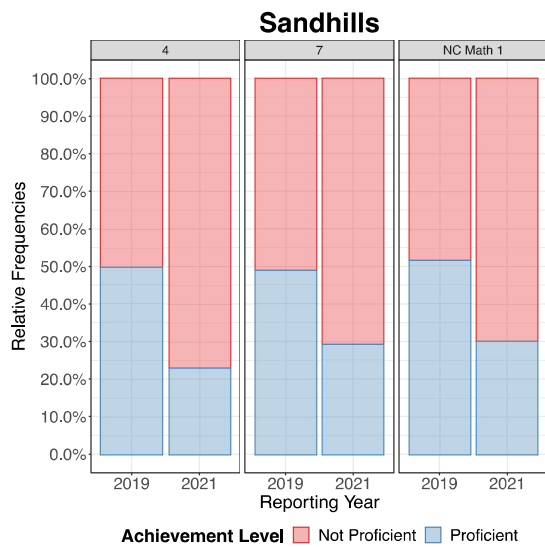
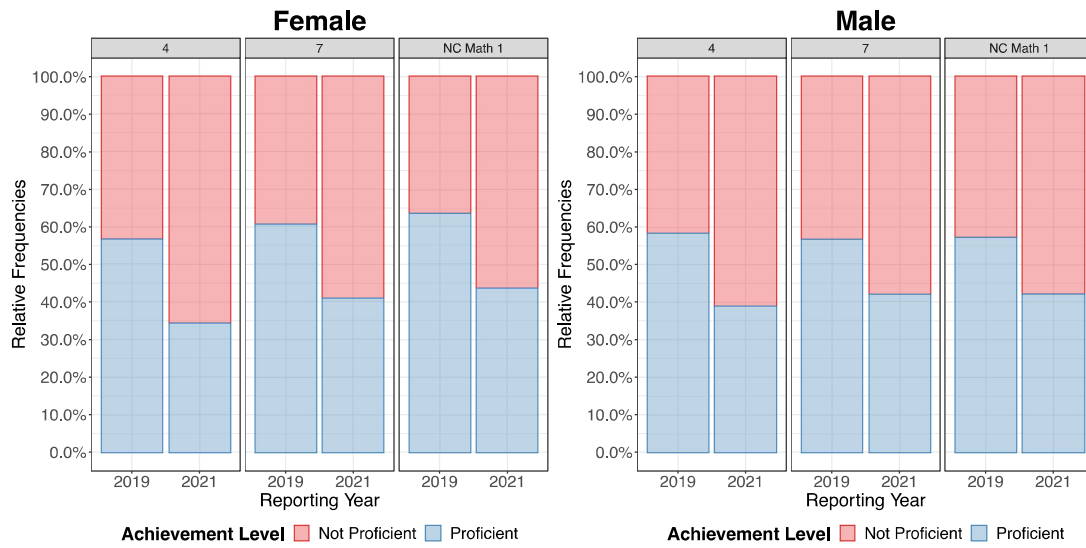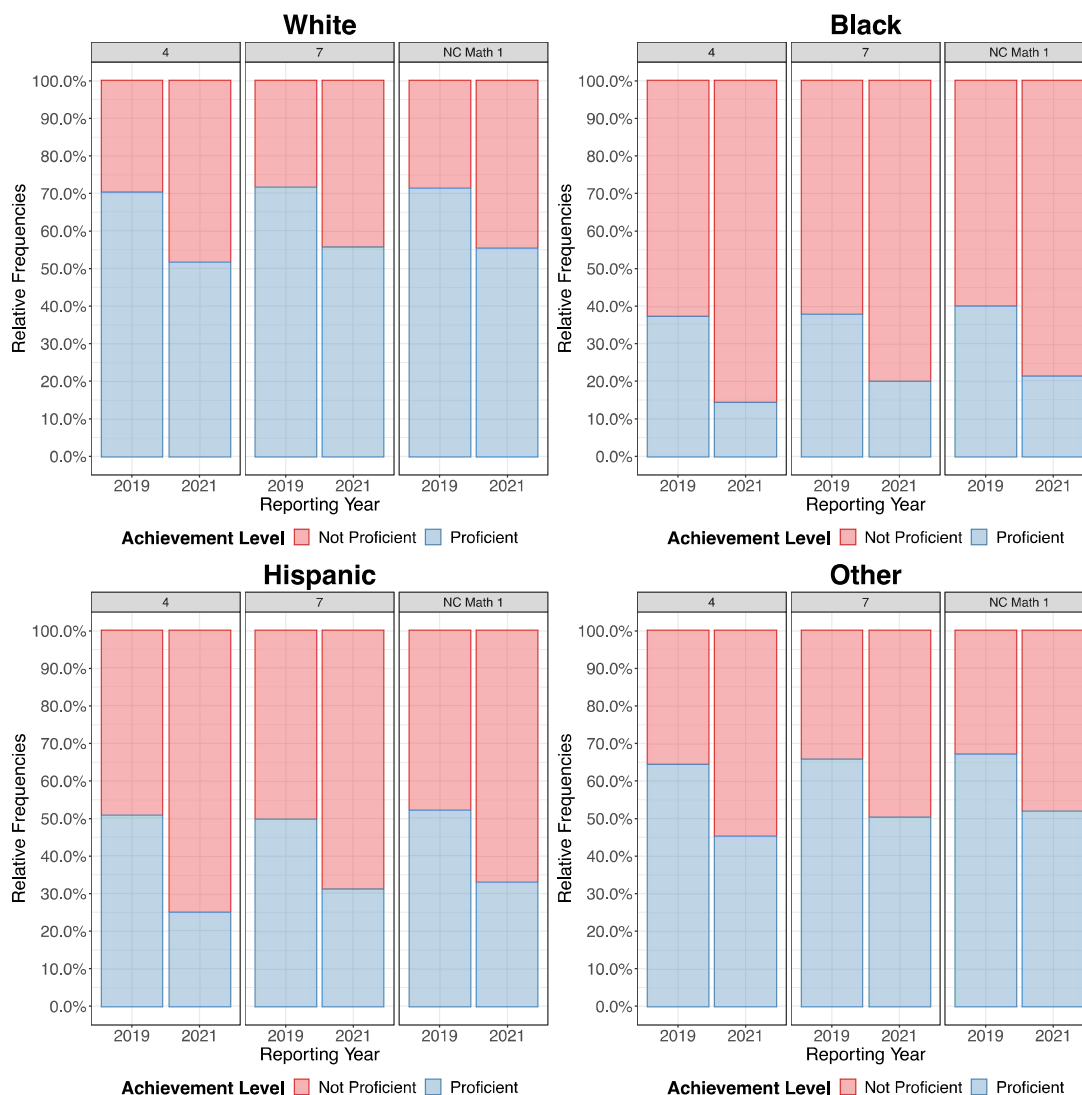Figure 6. Sex: Percentage of Students Classified as Grade-Level Proficient for the Two EOGs and NC Math 1

Figure 7. Ethnicity: Percentage of Students Classified as Grade-Level Proficient for the Two EOGs and NC Math 1

Item and Test Analyses

This section provides results for the comparison of the 2019 and 2021 student populations with respect to their performance on the test items and the overall test. The comparison was made using both classical test theory (CTT) and item response theory (IRT), which are briefly described below, for the overall population.

Classical Item Statistics

Proportion of Correct Responses. In CTT, item difficulty is measured by the proportion of correct

responses, which is calculated by dividing the number of students answering the item correctly by the total number of students (this proportion is often referred to as p-value). Accordingly, higher p-values indicate easier items, and conversely, lower p-values indicate more difficult items. Figures 8 to 10 show scatterplots of the p-values, by form for the two EOGs and NC Math 1, plotting 2019 p-values on the horizontal axis and 2021 p-values on the vertical axis. If the items were equally difficult for the 2019 and 2021 student populations, all points would lie near the identity line. However, all points were below the identity line, indicating that items had lower p-values for the 2021 administration compared to the 2019 administration; that is, the items were more difficult for the 2021 student population. This implies that the student population from 2021 was less proficient than the population from 2019.

Figure 8. Grade 4: Proportions of Correct Responses for the 2019 and 2021 Student Populations
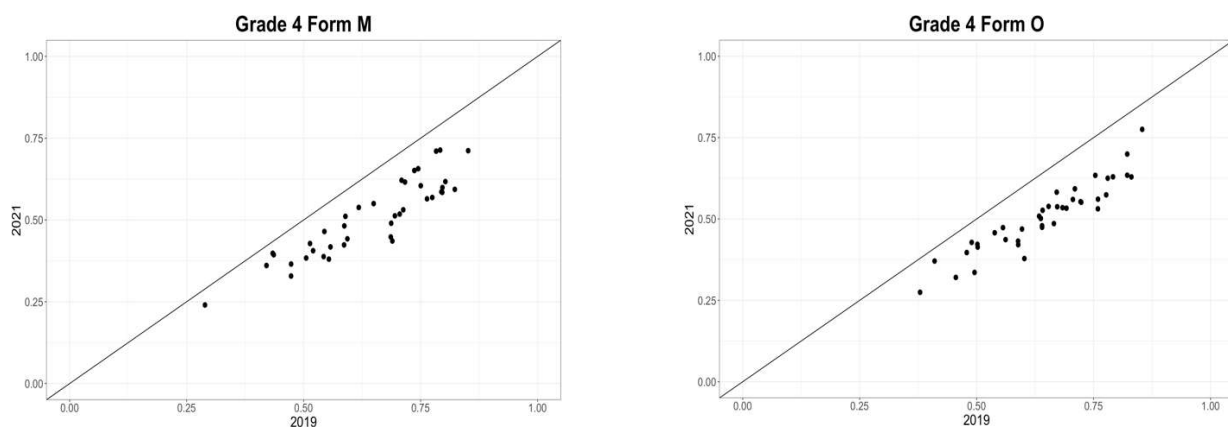
Figure 9. Grade 7: Proportions of Correct Responses for the 2019 and 2021 Student Populations
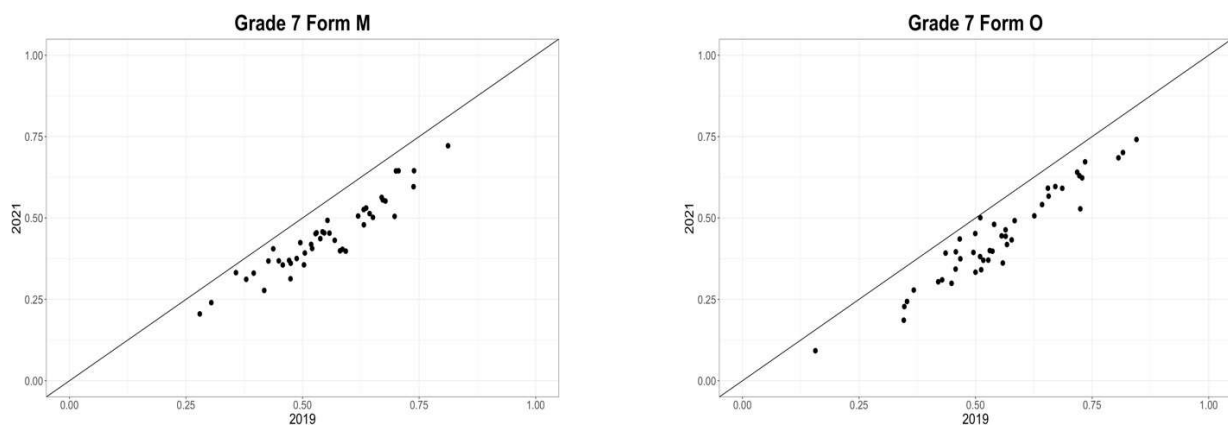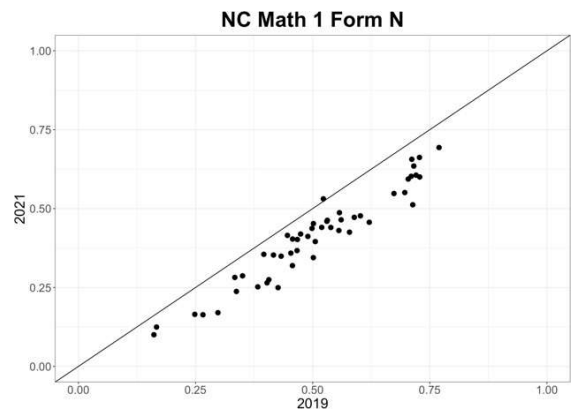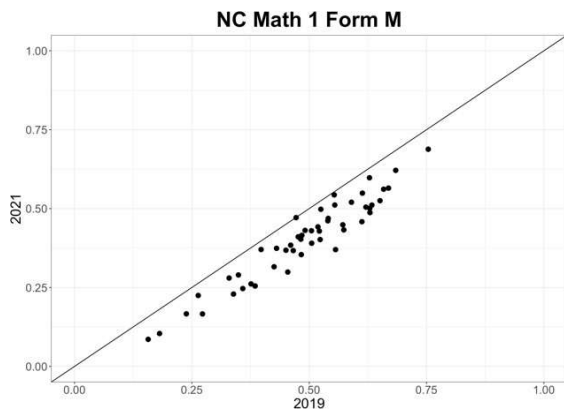

Figure 10. NC Math 1: Proportions of Correct Responses for the 2019 and 2021 Student Populations

NC Math 1 Form M / **NC Math 1 Form N**

IRT Parameters. For dichotomous items, IRT relates a person's estimated ability to the probability of correctly responding to an item using a mathematical model. In IRT, the higher an individual's estimated ability, the higher the probability of correct response. Theoretically, ability ranges from $-\infty$ and ∞; however, ranges such as (-3, 3) and (-4, 4) are often used in applications. In the mathematical model, an item is characterized by its discriminating power, difficulty, and guessingprobability. Item discrimination determines the rate at which the probability of answering an item correctly changes over ability. Higher values of the discriminating parameter indicate better differentiation among individuals on the ability continuum. Item difficulty measures the difficulty of an item. However, unlike the p-value, larger values of item difficulty indicate more difficult items. Assuming that there is no guessing involved in solving an item, item difficulty is the ability level at which 50% of the respondents endorse the correct answer. Item guessing accounts for guessing on an item and is defined as the probability of correctly responding to an item as the ability approaches $-\infty$.

Figures 11 to 13 display scatterplots for the item difficulty parameter estimates, by form for the two EOGs and NC Math 1, plotting item parameter estimates for 2019 on the horizontal axis and those

for 2021 on the vertical axis. The item difficulty parameter estimates were above the identity line. This suggests that items were more difficult for the 2021 student population compared to the 2019 student population; in other words,the student population from 2021 was less proficient than the student population from 2019. The results based on the IRT difficulty parameters were consistent with the results observed with the CTT p-values.

Figure 11. Grade 4: Item Difficulty Parameter Estimates for the 2019 and 2021 Student Populations
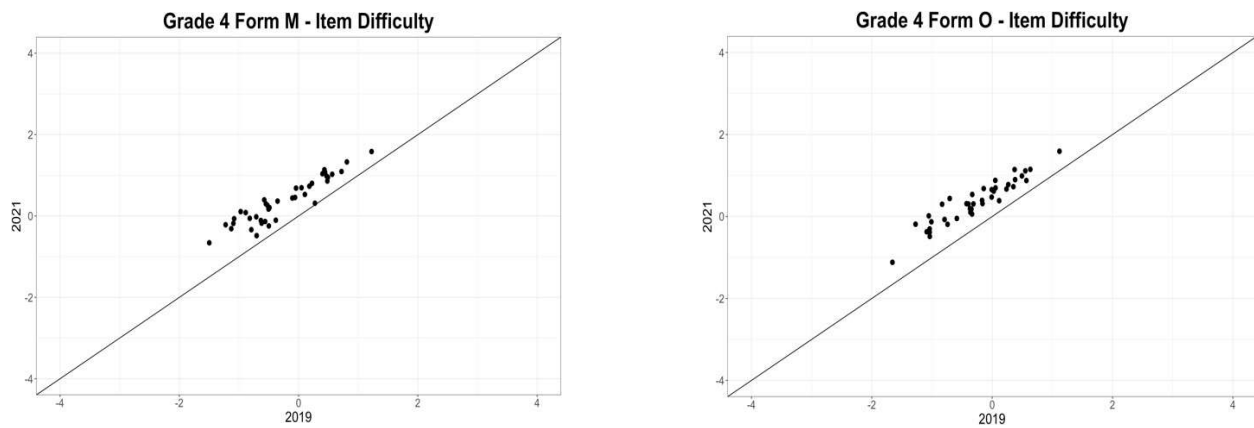
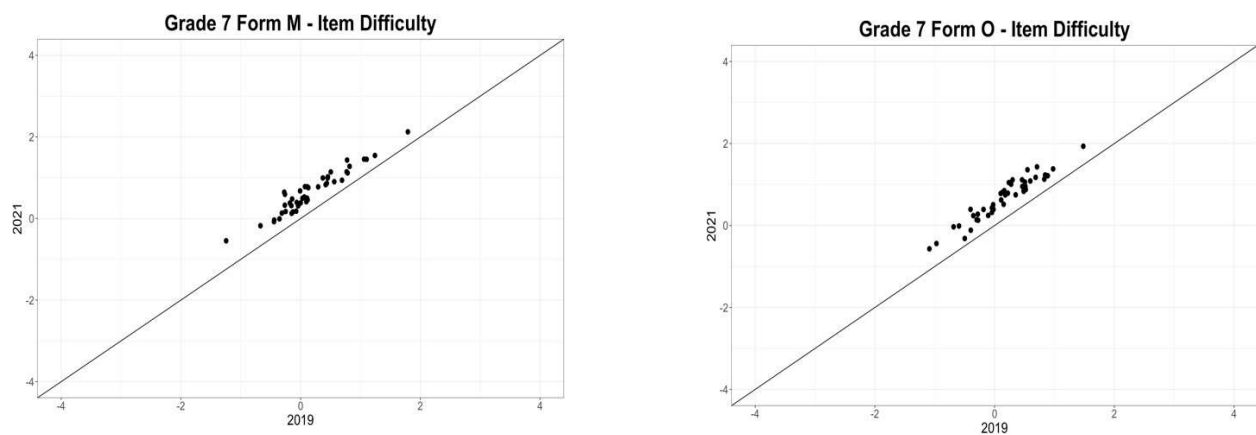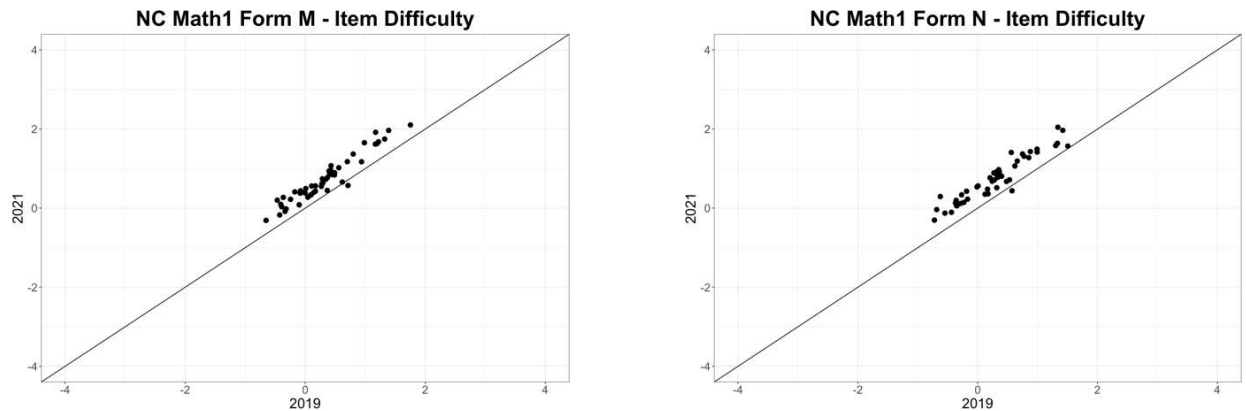Figure 12. Grade 7: Item Difficulty Parameter Estimates for the 2019 and 2021Student Populations

Figure 13. NC Math 1: Item Difficulty Parameter Estimates for the 2019 and 2021 Student Populations

Test Characteristic Curves. The mathematical model in IRT is often referred to as the item characteristic curve (ICC). The shape of the ICC is sigmoidal (i.e., an s-shaped curve) and provides the correct response probability at each ability level. The test characteristic curve (TCC) is the sum of ICCs over all items on a given test and provides the expected test score at each ability level. Figure 14 provides the TCCs of two test forms for each administration year. For the two EOGs and NC Math 1, the TCC for the 2021 administration was shifted to the right of the TCC for the 2019 administration. This indicates that students with the same ability have lower expected test scores for students from the 2021 population than the 2019 population. Stated differently, the test forms appeared to be more difficult for the 2021 student population, resulting in lower expected test scores. However, as grade level increased, the difference in expected test scores between the 2019 and 2021 student populations for each test form became smaller. Note that these findings were consistent with the findings observed with scale scores.

Figure 14. Test Characteristic Curves for the Two EOGs and NC Math 1
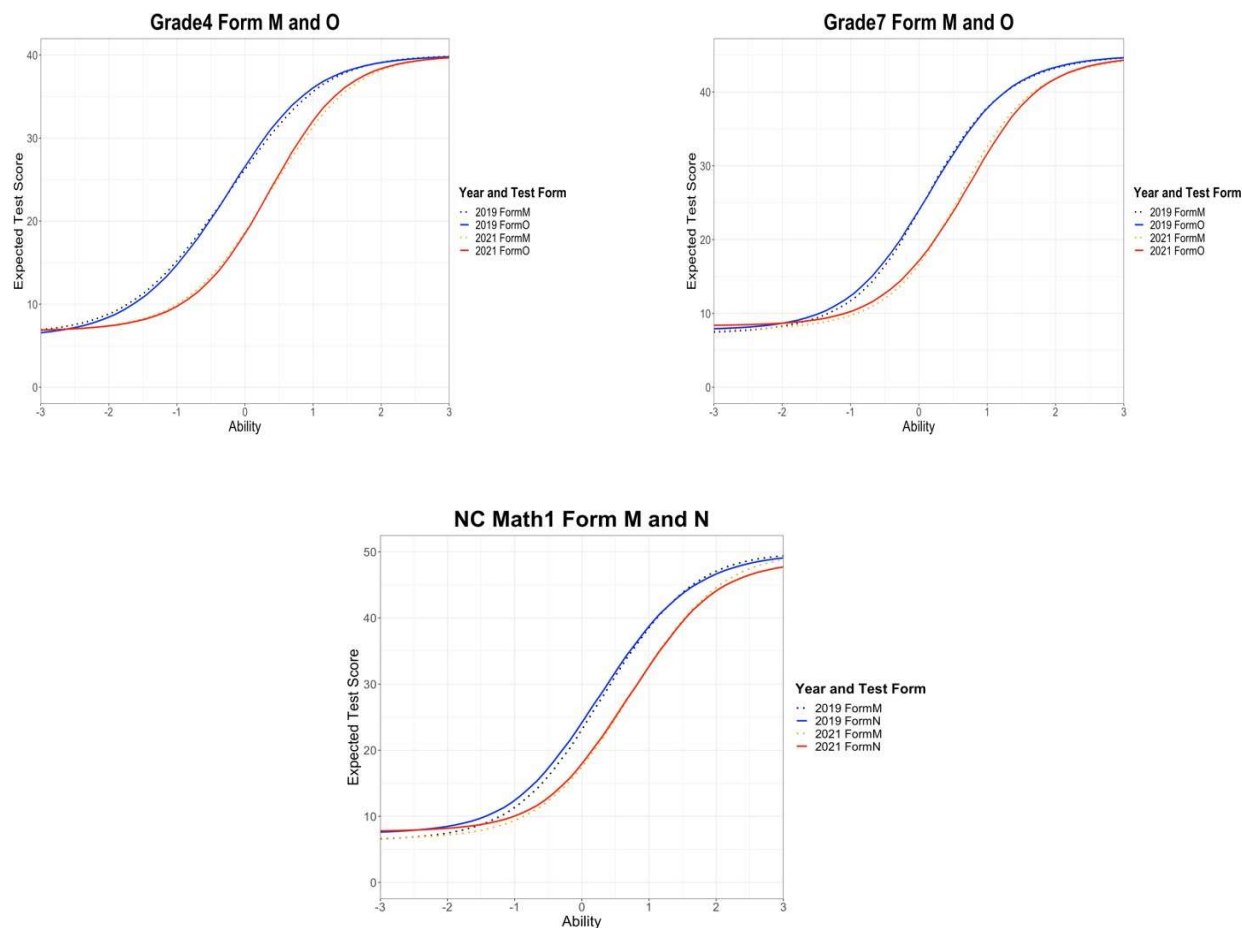
Ability Distributions. Figure 15 depicts the ability distributions of the 2019 and 2021 student populations for the two EOGs and NC Math 1. The dashed lines represent the mean ability for each year. For all three assessments, the ability distributions for the 2021 student populations were shifted to the left of the ability distributions for the 2019 student populations with a mean estimated ability of about -.5 for the 2021 student populations compared to mean estimated ability of 0 for the 2019 student populations, which indicates that the student populations from 2021 were less proficient than the student populations from 2019. In other words, on average, the abilities for the 2021student populations were about half standard deviation lower than the abilities for the 2019 student populations for the two EOGs and NC Math 1, a finding that was consistent with the result

observed with scale scores.

Figure 15. Ability Distributions for the two EOGs and NC Math 1
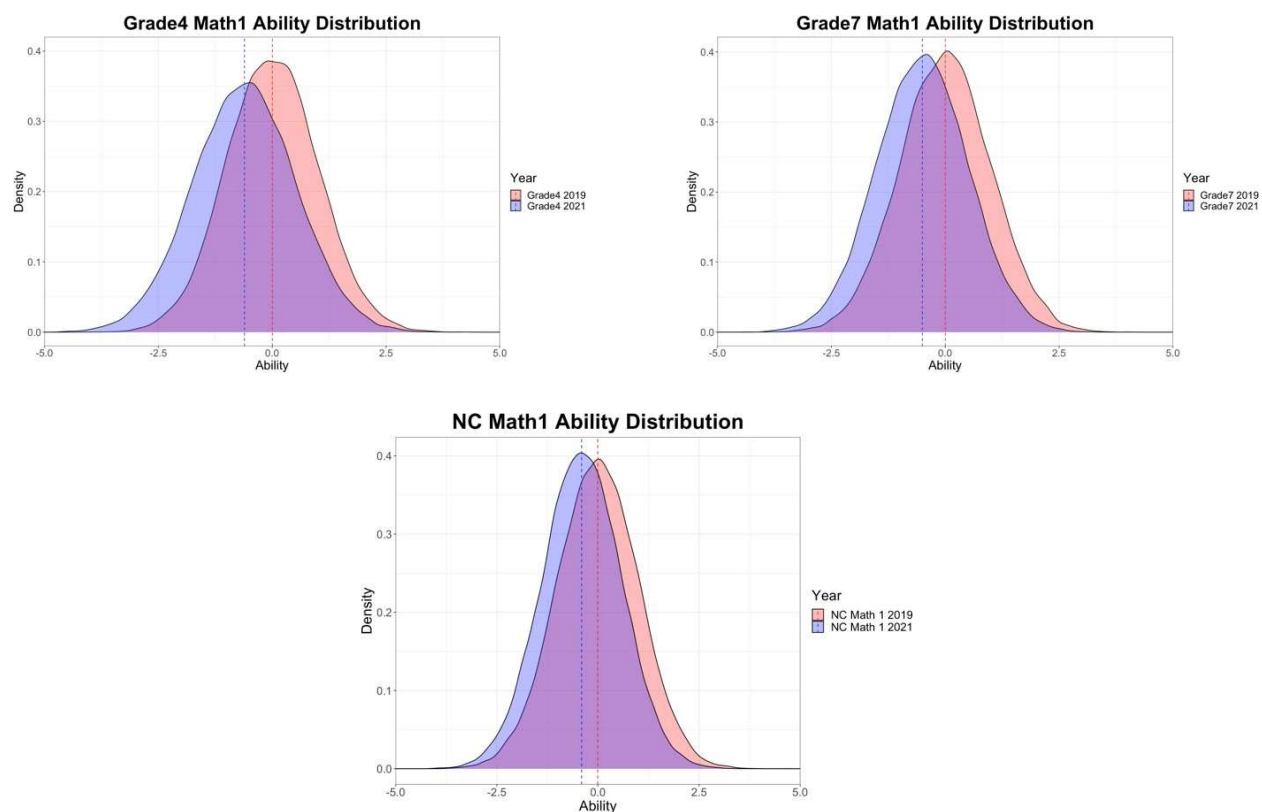
Summary

The results discussed above all indicate that the student performance from 2021 based on EOG mathematics Grades 4 and 7 and EOC NC Math 1 were lower than expected when compared to the performance of pre-COVID years from all PSU in North Carolina who took these same EOG and EOC assessments. In terms of both scale scores and IRT abilities, scores / abilities for the 2021 student populations were about half standard deviation lower than those for the 2019 student populations. In addition, the difference in performance was largest in Grade 4 and lessened slightly in Grade 7 and NC Math 1. Finally, the overall difference was generally consistent across subgroups, with the largest effects observed for Grade 4, in the Sandhills region, and among Black and Hispanic/Latino students.

There are two important caveats about the analyses and results presented in this study. First, although student participation on EOG and EOC assessments were close to the 95% criteria at the state level, enrollment data from 2021 indicate a sharp drop compared to historical trends. EOG and EOC participation across the state also varied by region and within subgroups. Therefore, there is no conclusive evidence that the decline in student enrollment was random and the 2021 cohort was similar to previous EOG and EOC cohorts. Second, the student experience from 2021 was very different across the state due to the outbreak and eventual risks associated to COVID-19, which caused disruptions to traditional instruction including various versions of hybrid and virtual learning opportunities to students.