



2018 No. 041

# TNReady Disruption Impact Study: End-of-Course and Grade-Level Results

## Final Report

**Prepared for:** Tennessee Department of Education  
Andrew Johnson Tower, 10th floor  
710 James Robertson Parkway  
Nashville, TN 37243  
Attn: Dr. Heather Peltier, Chief Assessment Officer  
Mary Batwalla, Assistant Commissioner

**Authors:** Matthew S. Swain  
Bethany H. Bynum  
Siqi (Lucy) Chen

**Prepared under:** Agency Tracking #: 33111 01318  
(HumRROS 18 35)

**Date:** July 27, 2018

# TNReady Disruption Impact Study: End-of-Course and Grade-Level Results

## Table of Contents

Key Findings .....	i
Executive Summary.....	ii
Background.....	4
End-of-Course Exams .....	5
Data Cleaning.....	5
Propensity Score Matching .....	6
Student-Level Analyses .....	8
Mean Exam Score Comparisons.....	8
2018 Exam Score Predictability.....	11
Examine Distributions of Predicted Student Scores.....	15
Compare Predictions of Disrupted Students to Non-Disrupted Students.....	15
Student-Level Summary.....	18
School-Level Analyses.....	19
School-Level Means .....	19
School-Level Classification.....	20
School-Level Summary .....	21
Invalidation.....	21
Invalidation Summary.....	23
Grade-Level Results.....	24
Data Cleaning.....	24
Propensity Score Matching .....	25
Student-Level Analyses .....	26
Mean Exam Score Comparisons.....	26
2018 Exam Score Predictability.....	29
Examine Distributions of Predicted Student Scores.....	34
Compare Predictions of Disrupted Students to Non-Disrupted Students.....	34
Student-Level Summary.....	37
School-Level Analyses.....	37
School-Level Means .....	37
School-Level Classification.....	38
School-Level Summary .....	39

## Table of Contents (continued)

Invalidation.....	39
Invalidation Summary.....	40
General Conclusions.....	41
References.....	42

## Key Findings

- During the 2018 TNReady testing window, across the end-of-course (EOC) and grade-level exams, 4% to 30% of students testing online experienced a disruption.
- Across all 28 EOC and grade-level exams, students who were involuntarily signed out during the test session and had to re-initiate the sign-in process, on average, scored lower than their non-disrupted peers. Thirteen of the 28 exams were not statistically significant at  $p < .05$ .
  - For the 12 EOC exams, disrupted students scored, on average, 4.6 scale score points lower than their non-disrupted peers for Chemistry and Biology exams (score range 500-900) and 2.2 scale score points lower for all other subjects (score range 250-400). Three of the 12 EOC exams were not statistically significant at  $p < .05$ .
  - For the 16 grade-level exams, disrupted students scored, on average, 7.2 scale score points lower for Science exams (score range 600-900) and 3.1 scale score points lower for all other subjects (score range 250-400). Ten of the 16 grade-level exams were not statistically significant at  $p < .05$ .
- Scores for students who experienced a lapse of over four hours between initial test sign-in and submission scored lower, on average, for eight of the 12 EOC exams and seven of the 16 grade-level exams, and were higher, on average, for four of the EOC exams and nine of the grade-level exams. Nineteen of the 28 exams were not statistically significant at  $p < .05$ .
  - For the EOC exams, disrupted students scored, on average, 4.6 scale score points lower for Chemistry and 1.7 scale score points lower on the other seven exams. Disrupted students scored, on average, 0.8 scale score points higher on Biology and 1.4 scale score points higher for the other three exams. Eleven of the 12 EOC exams were not statistically significant at  $p < .05$ .
  - For the grade-level exams, disrupted students scored, on average, 3.7 scale score points lower on three Science exams and 1.8 scale score points lower on four other exams. Disrupted students scored, on average, 7.9 scale score points higher on grade 5 Science and 2.3 scale score points higher on the other eight exams. Eight of these 16 exams were not statistically significant at  $p < .05$ .
- Scores for students whose experienced a disruption that resulted in a computer cache recovery request scored lower for five of the eight EOC exams and 11 of the 12 grade-level exams, and higher, on average for three of the EOC exams and one of the grade-level exams. Eighteen of the 20 exams were not statistically significant at  $p < .05$ .
  - For the EOC exams, disrupted students scored, on average, 6.0 scale score points lower on the Chemistry exam and 1.7 scale score points lower on four other exams. Disrupted students scored, on average, 1.0 scale score point higher for three exams. Seven of the eight EOC exams were not statistically significant at  $p < .05$ .
  - For the grade-level exams, disrupted students scored, on average, 6.3 scale score points lower on all three Science exams compared. For all other subjects, students scored, on average, 3.0 scale score points lower except for grade 7 Math where disrupted students scored 0.3 scale score points higher. Eleven of the 12 grade-level exams were not statistically significant at  $p < .05$ .
- After removing students who experienced a disruption, school means increased on average by 0.63 scale score points and 0.6% more students were considered “proficient.” For grade-level exams, after removing students who experienced a disruption, school means increased by 1.14 scale score points and 1.1% more students were considered “proficient.”

## Executive Summary

Several events occurred that disrupted the testing experiences for 4% to 30% of students who tested online during the 2018 TNReady testing window. Specifically, three types of disruptions occurred: (a) students were involuntarily signed out during the test session and had to re-initiate the sign-in process at least once (i.e., multiple sign-ins), (b) students lost connectivity or were booted off of the system and signed in again at a later time or date to finish their exam, resulting in over four hours elapsing between initial sign-in and test submission, and (c) due to system errors, a student's data was lost during the test session and the test administrator had to request recovery from the computer cache. Because of the wide-spread and systematic nature of these computer disruptions, the Tennessee Department of Education (TNDOE) contracted with the Human Resources Research Organization (HumRRO) to investigate the impact of these disruptions on students' test scores. This report describes the analyses that were conducted to determine the impact, if any, that computer disruptions had on students' test scores.

The foundation of our investigation is that students' test scores tend to exhibit consistency over time. That is, students who earn higher scores on a standardized test in one year tend to earn higher scores on the same or a highly similar standardized test in the next year. Therefore, we can use indicators of performance in 2017 to estimate what students' scores would be in 2018. However, it is also well documented that test scores for any individual student on any given day can also be impacted by factors other than his/her underlying knowledge or ability. One such factor could be a computer disruption, but there are many others. For example, a student might perform more poorly than expected on a given day due to things occurring in his or her home life, or because he/she was distracted, or not feeling well on the day of testing. Therefore, this investigation cannot definitively conclude that an *individual* student's performance was specifically impacted by a computer disruption rather than some other event in the student's life that was also present on the day of testing. The goal of this study was to determine if there are *trends* in the data that suggest a computer disruption had a systematic impact on student performance.

Because the computer disruptions did not affect most students, we used a set of variables to match disrupted and non-disrupted students to help estimate the impact of the disruption on students' test scores. By matching the samples on variables that are likely to predict students' scores, any difference between the two samples can be better attributed to the computer disruptions. We used several analyses to examine differences in scores between students who were disrupted and those who were not disrupted. Additionally, we investigated the impact of computer disruption on school means by considering alternative ways to compute school-level means, taking the disruptions into consideration. Finally, we examined the consequences of test administrator invalidation: whether these invalidated student records affected aggregated scores or were comprised of individuals who differed from the validated sample on several demographic variables.

The results of these analyses show a small, but consistent negative effect of computer disruption on students' test scores, particularly for the multiple sign-in disruption. On average, across all 28 end-of-course (EOC) and grade-level exams, disrupted students who had to re-initiate a sign-in multiple times earned lower scores than their non-disrupted peers and scored lower than expected. The analyses also suggest that the impact was not large, with score differences ranging from less than one to six scale score points for EOC exams and from less than one point up to 10 points for grade-level exams. This effect was also observed in the school-level aggregate means and more pronounced for grade-level exams. Because we ruled

out many other possible explanations for the difference by using propensity score matching, it is highly likely that the difference is due to the computer disruptions that occurred during the 2018 testing window. Due to this systematic negative effect, the test scores for students who experienced multiple sign-in attempts should be interpreted with this in mind.

For the students who experienced a disruption that resulted in over four hours elapsing between initial sign-in and test submission or for those that experience a disruption that resulted in a cache recovery request, the results were mixed. Although we observed lower scores for the disrupted sample on 30 of the 48 (62.5%) EOC and grade-level exams for the over four hours and cache recovery disruptions, the effect was not consistently detrimental, but at times were beneficial, with the disrupted sample scoring higher than the non-disrupted sample (18 of the 48 exams, or 37.5%). Additionally, for 14 of the 76 (18.4%) EOC and grade-level exams across all disruption types, the differences were less than one score point, generally suggesting little to no difference between the two samples.

Because of these disruptions or other irregularities during the test administration, test administrators could invalidate the 2018 test score for a given student (or group of students) thereby deleting their score from the record. Test administrators did not, however, invalidate the scores for every student who experienced a computer disruption and had discretion in determining which students experienced irregular administrations. Because of the computer disruption, the TNDOE expected a high number of records to be invalidated in 2018 and were concerned that the students whose scores were invalidated may not be representative of the state population. Invalidation of 2018 scores was rare (about 1.7% of students across EOCs, on average, and at most 4.1% and even rarer for grade-level exams at < 1.0%). The demographic characteristics (i.e., gender, race, economically disadvantaged, English language learner, and special education) of students with invalidated scores were similar to students with validated scores. This suggests that invalidation records were generally representative of the state population and were not associated with specific student characteristics.

# TNReady Disruption Impact Study: End-of-Course and Grade-Level Results

## Background

Several events occurred that disrupted the testing experiences for some students who tested online during the 2018 TNReady testing window. The TNReady includes 12 EOC exams and six grade-level online exams (grades 5 through 8) that are comprised of one, three, or four subparts, and a student could experience a disruption within each subpart of an EOC exam. The Tennessee Department of Education (TNDOE) wished to examine the impact of these disruptions on student performance. It contracted with the Human Resources Research Organization (HumRRO) to carry out this investigation. This report describes the analyses that were conducted to determine the impact, if any, that computer disruptions had on students' test scores. Specifically, they were interested in the following questions:

- To what extent did computer-based disruptions in online testing affect student and school scores?
- If there is evidence of an effect, how were student and school scores affected by disruptions?
- What does the analysis suggest for any guidance needed for interpretation and use of student and school scores?

During the 2018 testing window, the following issues occurred for at least some students who tested online: (a) students were involuntarily signed out during the test session and had to re-initiate the sign-in process at least once, (b) students lost connectivity or were booted off of the system and signed in again at a later time or date to finish their exam, resulting in over four hours elapsing between initial sign-in and test submission, and (c) due to system errors, a student's data was lost during the test session and the test administrator had to request recovery from the computer cache. Hereafter these disruptions are referred to as "multiple sign-in," "over four hours," and "cache recovery," respectively.

Because of these or other irregularities during the test administration, test administrators (TA) could invalidate the 2018 test score for a given student (or group of students) thereby deleting their score from the record. TAs did not, however, invalidate the scores for every student who experienced a computer disruption and had discretion in determining which students experienced irregular administrations. Because of the computer disruption, the TNDOE expected a higher number of records to be invalidated in 2018 than in previous years and were concerned that a high volume of invalidated student records could impact aggregate-level (school, district, and state) test scores. Accordingly, HumRRO was asked to examine the impact that the computer disruption had on student scores, as well as the impact that invalidated scores had on school, district, and state-level score means.

We present the data cleaning, procedures, and results first for the EOC exams and then for the grade-level exams.

## End-of-Course Exams

### Data Cleaning

The TN DOE provided 499,735 student records with scores across the 12 2018 EOC exams and indicators for each of the three disruption types<sup>1</sup> on each subpart of the EOC exams. The TN DOE had already performed some data screening, for example, removing home-schooled students from the data set. We were also provided with 2017 test scores for EOC exams and grades 3 through 8 English Language Arts (ELA), Math, and Science test scores. Prior year test scores are a key component in matching disrupted students to non-disrupted students because they are the best indicator of future test performance, particularly when the content is similar. Unlike grades 3 through 8, where most students take the same test as they progress through the grades, not all high school students take the same pattern of courses. Therefore, in finding the best prior year test score, our goal was to maximize sample size but also maintain theoretically related content across years. For example, when identifying the best matching student sample for Integrated Math I scores in 2018, we chose students who had 2017 Math scores (from grade 8) because this was the 2017 score with the highest sample size *and* the most theoretically consistent content. Table 1 lists the 2017 exam scores that were selected as the most relevant comparison for each 2018 EOC exam.

We continued the data cleaning process by removing student records that were missing key variables required to merge data files and to create matched pairs of students, including prior year test scores (EOC or grades 3 through 8), current year EOC scores, student ID, district ID, school ID, sex, race, grade, English Language Learners (ELL) status, economically disadvantaged status, and Special Education status. After cleaning the data, 446,620 records remained with complete data of the 12 EOC exams.

Table 1 also contains the total student sample size (total  $n$  with scores), the student sample size with complete data (2018 & 2017 Merged  $n$ ), and the student sample size experiencing each of the three computer disruption types. Some students experienced more than one disruption type and were included in each sample. Because we expected students to experience the three types of computer disruptions differently, we conducted analyses separately for each type of disruption and each of the EOC exams. If no effects were found, we planned to investigate the combined effects of two or more disruptions. However, we were concerned that a limited definition of disrupted students would lower sample size enough to make statistical estimation tenuous.

For the multiple sign-in disruption, we identified students as “disrupted” if they experienced a sign-in disruption at least once on more than one subpart of the exam. Most EOC exams have more than one subpart, so the decision rule is, essentially, at least one extra sign-in attempt on at least two subparts. For Biology and Chemistry, which have only one subpart each, a student was considered disrupted if he/she had one extra sign-in disruption on the only exam subpart. We chose to narrow this disruption because a relatively large percentage of students did experience one instance of needing to sign-in again on at least one subpart (more than 20%), and we could not find enough strong matches in the non-disrupted sample for all of these students. Furthermore, we reasoned that if we did not find an impact on test scores for students who were required to sign-in “two or more times,” then there would not be an impact for students who experienced the sign-in disruption only once. Therefore, for the purposes of our

---

<sup>1</sup> TN DOE request the primary test vendor identify students that experienced the three disruption types.



analyses, students with a multiple sign-in disruption on only one subpart of an exam with three or four subparts, were excluded from the disrupted and the non-disrupted samples.

For the over four hours disruption, students were considered disrupted if the time between their initial sign-in and their test submission was over four hours during one EOC subpart. For the cache recovery disruption, students were considered disrupted if a test administrator had to request a student’s test record be recovered from the computer cache on at least one EOC subpart. Because fewer than 100 students experienced the cache recovery disruption on the Biology, and Integrated Math I, II, and III exams, we excluded these EOC exams for the cache recovery disruption evaluation. We did this because any statistical comparisons based on such small sample sizes would be tenuous and potentially misleading.

**Table 1. Merged Total Sample Sizes and Disruption Type Sample Sizes by 2018 EOC Exam**

2018 EOC Exam	Total <i>n</i> with Scores	Most Relevant 2017 Exam	2018 & 2017 Merged <i>n</i>	Multiple Sign-in <sup>a</sup>		Over Four Hours		Cache Recovery <sup>b</sup>	
				<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Algebra I	50,566	Gr. 8 Math	45,982	2,695	5.86%	3,758	8.17%	544	1.18%
Algebra II	42,255	Geometry	33,428	1,350	4.04%	1,941	5.81%	681	2.04%
Biology	52,808	English I	24,153	3,567	14.77%	856	3.54%	--	--
Chemistry	42,859	Biology	31,524	3,210	10.18%	707	2.24%	712	2.26%
English I	54,298	Gr. 8 ELA	52,526	5,755	10.96%	10,473	19.94%	741	1.41%
English II	50,489	English I	49,508	4,879	9.85%	8,054	16.27%	201	0.41%
English III	40,666	English II	39,646	3,544	8.94%	5,352	13.50%	1,027	2.59%
Geometry	42,834	Algebra I	33,148	2,105	6.35%	3,523	10.63%	206	0.62%
Int. Math I	11,942	Gr. 8 Math	10,836	506	4.67%	891	8.22%	--	--
Int. Math II	11,404	Int. Math I	10,441	378	3.62%	647	6.20%	--	--
Int. Math III	7,339	Int. Math II	6,807	274	4.03%	216	3.17%	--	--
U.S. History	39,160	English II	33,781	2,032	6.02%	3,594	10.64%	969	2.87%

Note. Int. Math = Integrated Math. Gr. 8 = Grade 8. ELA = English Language Arts.

<sup>a</sup>Disruption was defined as experiencing at least one multiple sign-in attempt on more than one subpart of the exam. For the 2018 Biology and Chemistry EOC exams, which each have one subpart, students with just one sign-in disruption were included in the disrupted sample.

<sup>b</sup>Excluded EOC exams for which the sample size was too small ( $n < 100$ ).

### **Propensity Score Matching**

Propensity matching is a statistical approach used to match samples on a set of variables that are likely to be related to the outcome of interest when random group assignment is not possible. This type of procedure is appropriate for this study because students were not randomly assigned to be in the disrupted group. The objective of propensity score matching was to match students who were disrupted with students who were not disrupted on variables that contributed to 2018 test scores, such as demographic variables and prior year scores. If the samples are effectively matched, then any observed differences between the two samples on 2018 test scores are more likely to be due to the disruption. For each disruption type, the result was two samples of students (disrupted and not-disrupted) who were as closely matched as possible, except for their experience with computer disruptions and 2018 EOC exam scores.

Students in the disrupted samples were matched with non-disrupted students on<sup>2</sup>:

- Race
- Sex
- Student grade-level
- Relevant prior test scores
- English Language Learner status
- Special Education status
- Economically disadvantage status
- School-level achievement
- School-level proportion of economically disadvantage students

Prior to propensity score matching, we examined differences between the disrupted and non-disrupted samples to describe the need for matching. We use a standardized effect size, Cohen's  $d$ , to make comparisons<sup>3</sup>. There were several moderate to large differences between the disrupted and non-disrupted samples prior to matching with Cohen's  $d = .77$ ,  $.55$ , and  $1.22$ , for multiple sign-in, over four hours, and cache recovery, respectively. These differences were for the proportion of Algebra I Black Students, the proportion of Integrated Math II students who were economically disadvantaged, and school-level prior test scores for Geometry students, respectively. These results suggest that propensity matching is necessary to ensure the two samples are equivalent.

To determine the matched samples, we first used the matching variables listed above to predict the probability of being disrupted by disruption type. We did this using logistic regression and regressed group membership (disrupted or non-disrupted) onto the matching variables. The extent to which logistic regression can "explain" the dichotomous outcome was evaluated using a pseudo  $R^2$  index, which can be interpreted like a  $R^2$  value in multiple regression. The pseudo  $R^2$  values of the logistic regression were small, ranging from  $.0006$  to  $.0668$ . The small values suggest that overall, the combination of prior year student achievement, demographics, grade and English Language Learner, special education, and economic disadvantage status had little relationship to the likelihood that a student experienced disruption. From these three logistic regression equations, we saved the likelihood or probability that each student was disrupted. These predicted probabilities summarize a student's *profile* on the prediction variables. That is, two students with the same values on all matching variables listed above will have the same predicted probability.

Next, the predicted probability for each student in the disrupted sample was matched to the student with the closest predicted probability in the non-disrupted sample. The sampling was done without replacement so that each student in the disrupted sample was matched with a unique student in the non-disrupted sample. The largest difference between the predicted probabilities was  $.0134$ . This value is well within the maximum difference of  $.20$  that has been shown to reduce bias and produce accurate group difference estimates (Austin, 2009; Connelly,

---

<sup>2</sup> A preliminary report did not include school-level variables; thus, the matched samples and mean results will differ.

<sup>3</sup> We used Cohen's  $d$  to compare proportions of dichotomous variables (i.e., gender, race) to assess for balance between the two samples. We also provide  $t$ -tests to compare all demographic variables and prior year test performance.

Sackett, & Waters, 2013). The results suggest that every student in the disrupted sample was matched with a student in the non-disrupted sample that had a nearly identical predicted disruption probability.

To further evaluate the similarity of the matched samples for each disruption type, we examined the mean difference between samples on the matching variables. There was very little difference between the samples. The mean effect size (Cohen's  $d$ ) across all matching variables was .00 (range = -.08 to .08), .00 (range = -.11 to .13), and .01 (range = -.26 to .16) for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. Cohen's  $d$  effect sizes near zero suggest that the samples were effectively balanced on the matching variables. The three largest differences between the samples, prior to matching, were reduced to -.01, .04, and .01 after matching. This suggests that the matching was successful even on the variables with the largest differences in proportion prior to matching.

We did find one effect size greater than  $|.20|$  between disrupted and non-disrupted economically disadvantage students on the Geometry EOC exam for cache recovery disruption. However, the difference in economically disadvantaged students between the matched disrupted and non-disrupted samples was less than 3%. A summary of the means, standard deviations, and effect sizes before and after matching are found in Appendix A.<sup>4</sup>

### ***Student-Level Analyses***

Using the matched samples for each disruption type, we examined whether test scores of students who experienced computer disruptions differed from the test scores of students who were not disrupted. By matching the samples on several variables known to be related to student exam scores, we controlled for the impact of these matching variables on group differences. Therefore, any observed differences between the two samples are more likely to be due to computer disruptions. We used several analyses to examine differences in 2018 EOC scale scores.

#### ***Mean Exam Score Comparisons***

Below we summarize the mean EOC scale score differences between the disrupted and matched non-disrupted sample. If computer disruptions had no overall impact on student test performance, then the mean scores of the 2018 exams should be very similar. On the other hand, observed differences in mean 2018 EOC scores provides evidence that computer disruption did impact test performance. Comparisons of the 2018 EOC scores are presented in Tables 2, 3, and 4 for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. Means and standard deviations (SD) were compared using an independent-samples  $t$ -test ( $t$ -value) and Cohen's  $d$  effect size estimates. Due to the large sample sizes, small mean differences may result in statistically significant results but have little practical significance. Therefore, we recommend focusing on the Cohen's  $d$  values as a standardized measure of practical significance. Cohen's  $d$  values are interpreted as the difference in standard deviations between the two samples. As rules of thumb, Cohen suggested .20, .50, and .80 as small, medium, and large effects, respectively (Cohen, 1992). However, scale score differences may be more meaningful so these are discussed in text for the largest Cohen's  $d$  values. The

---

<sup>4</sup> The Appendices are under separate cover in Volume II: Appendices. A copy may be obtained by contacting [mswain@humrro.org](mailto:mswain@humrro.org).

range of EOC scale scores is 200-450 for all exams except for Biology and Chemistry, which range from 500-900.

For the multiple sign-in disruption, the mean score for the disrupted sample was lower than the mean score for the matched non-disrupted matched sample for all 2018 EOC exams. However, these differences were small. All EOC exams exhibited small effect sizes, that is, a Cohen's *d* value < |.20|. The largest effect size difference was observed for Integrated Math II with a mean score difference of 4.5 scale score points and a Cohen's *d* value of -.16.

For the over four hours disruption, the mean score differences were also small with Cohen's *d* effect sizes ranging from -.11 to .10. Integrated Math I, II, and III showed small determinantal effects, however, these results were based on low sample sizes, making conclusions regarding possible effects tenuous. Additionally, the disrupted sample scored higher than the matched non-disrupted sample on four exams. Many of the differences were less than one score point, generally, suggesting little to no difference between the two samples.

For the cache recovery disruption, all effect sizes were small, ranging from -.19 to .10. The results showed a mixed effect with the disrupted sample scoring lower than the non-disrupted sample for some exams, but higher than the non-disrupted sample on other exams. U.S. History showed the largest effect size, with mean score difference of 3.20 scale score points.

**Table 2. Mean EOC Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Multiple Sign-In Disruption Type**

2018 EOC Exam	Disrupted		Non-Disrupted		<i>t</i> -value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
Algebra I	2,695	292.13 (31.85)	2,695	294.77 (31.76)	-3.05**	-.08
Algebra II	1,350	289.13 (32.33)	1,350	291.17 (31.46)	-1.67	-.06
Biology	3,567	683.03 (44.46)	3,567	685.94 (44.03)	-2.78**	-.07
Chemistry	3,210	670.89 (86.43)	3,210	677.13 (86.52)	-2.89**	-.07
English I	5,755	319.09 (15.16)	5,755	320.20 (14.52)	-4.00***	-.07
English II	4,879	308.09 (13.03)	4,879	308.69 (13.41)	-2.24*	-.05
English III	3,544	318.78 (16.42)	3,544	320.50 (15.72)	-4.51***	-.11
Geometry	2,105	294.72 (31.91)	2,105	297.27 (31.66)	-2.60**	-.08
Integrated Math I	506	282.71 (38.58)	506	287.06 (37.79)	-1.81	-.11
Integrated Math II	378	283.43 (28.07)	378	287.93 (28.06)	-2.21*	-.16
Integrated Math III	274	278.28 (40.15)	274	279.84 (37.70)	-0.47	-.04
U.S. History	2,032	324.61 (19.38)	2,032	325.94 (19.48)	-2.19*	-.07

Note. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on more than one subpart of the exam (except for Biology and Chemistry, which have one subpart each).

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

**Table 3. Mean EOC Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Over Four Hours Disruption Type**

2018 EOC Exam	Disrupted		Non-Disrupted		t-value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
Algebra I	3,758	303.52 (30.64)	3,758	304.47 (29.51)	-1.38	-.03
Algebra II	1,941	299.61 (29.19)	1,941	296.63 (31.09)	3.08**	.10
Biology	856	693.09 (42.36)	856	692.34 (40.22)	0.38	.02
Chemistry	707	700.04 (83.63)	707	704.68 (89.54)	-1.01	-.05
English I	10,473	322.72 (14.66)	10,473	323.05 (14.16)	-1.66	-.02
English II	8,054	311.10 (12.86)	8,054	311.20 (12.67)	-0.49	-.01
English III	5,352	322.40 (15.91)	5,352	322.70 (15.17)	-0.97	-.02
Geometry	3,523	304.68 (30.09)	3,523	304.10 (30.60)	0.81	.02
Integrated Math I	854	295.39 (41.42)	854	299.04 (38.76)	-1.88	-.09
Integrated Math II	647	287.60 (30.47)	647	289.56 (27.55)	-1.21	-.07
Integrated Math III	216	286.87 (41.62)	216	291.21 (38.18)	-1.13	-.11
U.S. History	3,594	330.23 (19.31)	3,594	329.65 (19.00)	1.28	.03

Note. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on one subpart of the exam.  
 \**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

**Table 4. Mean EOC Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Cache Recovery Disruption Type**

2018 EOC Exam	Disrupted		Non-Disrupted		t-value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
Algebra I	544	297.87 (28.03)	544	299.63 (27.82)	-1.04	-.06
Algebra II	681	287.84 (32.01)	681	287.34 (33.94)	0.28	.02
Biology <sup>a</sup>	--	--	--	--	--	--
Chemistry	712	663.51 (75.18)	712	669.55 (75.96)	-1.51	-.08
English I	741	325.60 (15.61)	741	326.80 (14.07)	-1.56	-.08
English II	201	312.00 (11.56)	201	310.76 (12.21)	1.05	.10
English III	1,027	323.82 (14.92)	1,027	324.48 (13.84)	-1.04	-.05
Geometry	206	317.13 (26.28)	206	315.81 (29.04)	0.48	.05
Integrated Math I <sup>a</sup>	--	--	--	--	--	--
Integrated Math II <sup>a</sup>	--	--	--	--	--	--
Integrated Math III <sup>a</sup>	--	--	--	--	--	--
U.S. History	969	327.20 (16.98)	969	330.40 (17.04)	-4.14***	-.19

Note. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on one subpart of the exam.

<sup>a</sup>Group sample sizes were too small (*n* = 100) to present a stable mean for comparison.

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

## 2018 Exam Score Predictability

Another way to examine the potential impact of computer disruptions is to determine the extent to which disruption status helps explain variance in exam scores. Specifically, we examined if disruption added to the predictability of the 2018 test scores beyond other known indicators of performance. The covariates included in this model were prior year test score, sex, race, grade-level, prior year school-level test score, school-level economically disadvantaged status, and student-level indicators of English Language Learner, economically disadvantaged, or special education status. If a computer disruption impacted scores, then inclusion of the disruption indicator (0 = not disrupted, 1 = disrupted) in the multiple regression model should add to the estimation of 2018 scores, as indicated by the size of the multiple regression coefficient ( $R^2$ ). The addition of any variable to a regression model should increase the multiple regression coefficient ( $R^2$ ) but it is the degree of increase that indicates if that variable contributes to the prediction of the test score. We conducted regression modeling analyses separately for each EOC exam and each disruption type, where sample size permitted.

Table 5 shows the  $R^2$  values for each regression model with and without the disruption variable, the change in  $R^2$  between the two models ( $\Delta R^2$ ), and the unstandardized regression coefficient for disruption ( $b$ ).  $R^2$  values can be interpreted as the proportion of test score variance explained by the model. An  $R^2$  of .575 indicates 57.5% of the variance in Algebra I 2018 scores is explained by all variables mentioned above. After adding the group membership variable for disruption (0 = not disrupted, 1 = disrupted) to the model, the variance explained increases to .577 or a .002 change ( $\Delta R^2$ ). This  $\Delta R^2$  reflects a .2% increase in variance explained, which is small. The unstandardized coefficient for disruption ( $b$ ) is on the metric of the EOC scale scores and indicates the scale score difference between those not disrupted and those disrupted, after controlling for the other variables in the model. These values are statistically significant, as noted by the test of the regression coefficient ( $t$ -value). However, these coefficients can be statistically significant due to large sample sizes. Therefore, overall effects should be interpreted considering the range of EOC exam scores and the standardized effect size (in this case,  $\Delta R^2$ ) which, at most, accounted for an additional 1.6% of the variance in 2018 exam scores. Again, the range of EOC scale scores is 200-450 for all exams except for Biology and Chemistry, which range from 500-900.

In general, for all disruption types, there was very little or no change in the  $R^2$  values when the disruption variable was included in the regression model, indicating that disruption added very little to the prediction of 2018 scores beyond the other factors known to impact test scores. The direction of the unstandardized regression coefficients ( $b$ ) was generally negative meaning that disruption lowered observed scores, controlling for the other variables in the model. This result coincides with the individual-level mean differences.

**Table 5. Incremental Validity Estimation of Disruption by Disruption Type and EOC Exam**

Disruption Type	EOC Exam	n	Covariates		$\Delta R^2$	Disruption <i>b</i> estimate	<i>t</i> -value
			Covariates Only $R^2$	+ Disruption $R^2$			
Multiple Sign-In	Algebra I	5,390	.575	.577	.002	-2.835	-5.02***
	Algebra II	2,700	.354	.356	.002	-2.516	-2.54*
	Biology	7,134	.405	.406	.001	-2.698	-3.34***
	Chemistry	6,420	.436	.436	.001	-5.187	-3.19**
	English I	11,510	.574	.575	.001	-1.053	-5.83***
	English II	9,758	.611	.611	.001	-0.623	-3.73***
	English III	7,088	.488	.489	.002	-1.367	-5.00***
	Geometry	4,210	.436	.438	.001	-2.356	-3.19**
	Int. Math I	1,012	.627	.629	.002	-3.517	-2.38*
	Int. Math II	756	.397	.404	.008	-4.964	-3.10**
	Int. Math III	548	.436	.436	<.001	0.031	0.01
U.S. History	4,064	.448	.449	.001	-1.279	-2.82**	
Over Four Hours	Algebra I	7,516	.606	.606	<.001	-0.594	-1.36
	Algebra II	3,882	.424	.425	<.001	1.273	1.72
	Biology	1,712	.434	.434	<.001	-0.648	-0.43
	Chemistry	1,414	.496	.496	<.001	-1.506	-0.46
	English I	20,946	.604	.604	<.001	-0.275	-2.19*
	English II	16,108	.625	.625	<.001	-0.311	-2.52*
	English III	10,704	.520	.520	<.001	-0.419	-2.01*
	Geometry	7,046	.498	.498	<.001	-0.354	-0.69
	Int. Math I	1,708	.701	.702	.001	-2.293	-2.15*
	Int. Math II	1,294	.457	.460	.003	-3.211	-2.68**
	Int. Math III	432	.506	.508	.001	-2.940	-1.06
U.S. History	7,188	.499	.499	<.001	0.737	2.30*	
Cache Recovery	Algebra I	1,088	.542	.543	.001	-1.681	-1.46
	Algebra II	1,362	.310	.310	<.001	-0.264	-0.18
	Chemistry	1,424	.314	.315	.001	-5.330	-1.60
	English I	1,482	.638	.639	.001	-0.931	-1.99*
	English II	402	.661	.652	.002	-0.160	-0.22
	English III	2,054	.515	.526	.003	-1.337	-3.01**
	Geometry	412	.589	.564	<.001	0.531	0.29
	U.S. History	1,938	.423	.448	.016	-4.539	-7.74***

Note. Int. Math = Integrated Math. *t*-value = Test of the unstandardized regression coefficient for Disruption *b* estimate. Results for Biology, Integrated Math I, Integrated Math II, and Integrated Math III EOC exams are not presented for the cache recovery disruption due to low sample sizes in the disrupted sample. Covariates in the model were prior year test scores, sex, race, grade-level, and indicators for students who are English Language Learners, Economically Disadvantaged, or Special Education.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

Next, using all the indicators of test performance available, we estimated regression equations to predict 2018 test scores *separately* for each sample (disrupted and non-disrupted) for each EOC exam and disruption type. Whereas the preceding regression models analyzed the predictability of test scores with disrupted and non-disrupted students combined, we do not know the strength of the prediction for each sample separately. If students' performance was affected by the computer disruption, the strength of the prediction, as indicated by  $R^2$ , should be lower for the disrupted students than for the non-disrupted students. A lower  $R^2$  coefficient means that students' performance in the disrupted sample was not predicted as well as the non-disrupted sample. This is yet another way to gauge whether there was a general impact across students due to the computer disruptions.

Tables 6 through 8 present the  $R^2$  values for the disrupted and non-disrupted samples for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. Overall, for most EOC exams, the 2018 test scores were fairly well predicted for both disrupted and non-disrupted samples, with 31% to 62% of the variance accounted for by the available set of predictor variables. Generally, there were slightly higher  $R^2$  values for the non-disrupted samples, for all three disruption types, as shown by the positive  $R^2$  difference values. This means that 2018 exam scores were not quite as well predicted by the available set of prediction variables (or covariates), and suggests the computer disruption did, perhaps, have some impact on exam scores. The difference in variance accounted for ranged from  $< .1\%$  to  $10.1\%$  for all but one comparison, suggesting that the impact was small. For some EOC exams the  $R^2$  values were higher for the disrupted sample compared to the non-disrupted sample, with differences ranging from  $< .1\%$  to  $5.9\%$ . There was a 20% difference in variance explained between the disrupted and non-disrupted students on the Integrated Math II exam for the multiple sign-in disruption. This could be a statistical anomaly, or it could suggest that, for this particular EOC exam, the multiple sign-in disruption had a relatively larger impact on our ability to predict 2018 scores.

Across EOC exams, the  $R^2$  difference was generally larger for the non-disrupted sample than the disrupted sample. This difference could be due to differences in variance and not wholly attributable to differences in prediction. We computed Root Mean Square Error (RMSE) estimates to account for differences in sample variance. RMSE is the root of the variance of the residuals between observed and predicted scores. In other words, it shows us how well the model "fits" or replicates the observed scores. If the RMSE values are similar across samples, we have similar model fit between the samples. The RMSE differences were small (compared to the RMSE values), ranging from  $-4.4$  to  $4.4$ . This indicates that the model fit for the two samples was similar and for some EOC exams the model replicated the observed score better for the non-disrupted sample and for other EOC exams the model replicated the observed scores better for the disrupted sample.



**Table 6. Predictability of 2018 EOC Scale Scores for Non-Disrupted and Disrupted Groups – Multiple Sign-In Disruption Type**

Content	<i>n</i>	Non-Disrupted R <sup>2</sup>	Disrupted R <sup>2</sup>	R <sup>2</sup> Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
Algebra I	2,695	.594	.562	.032	20.29	21.13	-0.84
Algebra II	1,350	.338	.377	-.039	25.73	25.66	0.07
Biology	3,567	.423	.391	.032	33.51	34.77	-1.26
Chemistry	3,210	.445	.432	.013	64.60	65.29	-0.70
English I	5,755	.588	.563	.025	9.33	10.03	-0.70
English II	4,879	.621	.603	.018	8.27	8.23	0.05
English III	3,544	.517	.467	.050	10.95	12.02	-1.06
Geometry	2,105	.454	.424	.030	23.47	24.29	-0.82
Integrated Math I	506	.611	.658	-.047	23.90	22.88	1.03
Integrated Math II	378	.517	.316	.201	19.90	23.65	-3.74
Integrated Math III	274	.423	.482	-.059	29.41	29.67	-0.26
U.S. History	2,032	.462	.441	.022	14.33	14.55	-0.21

Note. R<sup>2</sup> Difference is the difference between the Non-Disrupted R<sup>2</sup> and the Disrupted R<sup>2</sup>.

**Table 7. Predictability of 2018 EOC Scale Scores for Non-Disrupted and Disrupted Groups – Over Four Hours Disruption Type**

Content	<i>n</i>	Non-Disrupted R <sup>2</sup>	Disrupted R <sup>2</sup>	R <sup>2</sup> Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
Algebra I	3,758	.611	.604	.007	18.44	19.31	-0.87
Algebra II	1,941	.430	.422	.008	23.56	22.28	1.28
Biology	856	.454	.424	.029	29.97	32.40	-2.43
Chemistry	707	.503	.505	-.002	63.83	59.47	4.36
English I	10,473	.609	.599	.010	8.86	9.28	-0.43
English II	8,054	.633	.619	.014	7.69	7.94	-0.26
English III	5,352	.553	.493	.060	10.16	11.35	-1.18
Geometry	3,523	.508	.490	.018	21.52	21.54	-0.01
Integrated Math I	854	.704	.701	.002	21.26	22.81	-1.55
Integrated Math II	647	.490	.452	.038	19.93	22.85	-2.92
Integrated Math III	216	.549	.485	.065	26.51	30.90	-4.40
U.S. History	3,594	.491	.510	-.020	13.59	13.54	0.05

Note. R<sup>2</sup> Difference is the difference between the Non-Disrupted R<sup>2</sup> and the Disrupted R<sup>2</sup>.

**Table 8. Predictability of 2018 EOC Scale Scores for Non-Disrupted and Disrupted Groups – Cache Recovery Disruption Type**

Content	<i>n</i>	Non-Disrupted R <sup>2</sup>	Disrupted R <sup>2</sup>	R <sup>2</sup> Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
Algebra I	544	.518	.574	-.056	19.56	18.53	1.03
Algebra II	681	.364	.263	.101	27.32	27.59	-0.26
Chemistry	712	.334	.304	.030	62.57	63.31	-0.74
English I	741	.634	.651	-.017	8.59	9.30	-0.71
English II	201	.679	.655	.024	7.14	6.99	0.14
English III	1,027	.546	.495	.051	9.39	10.68	-1.29
Geometry	206	.588	.647	-.059	19.15	16.02	3.13
U.S. History	969	.428	.453	-.025	12.96	12.63	0.33

Note. R<sup>2</sup> Difference is the difference between the Non-Disrupted R<sup>2</sup> and the Disrupted R<sup>2</sup>. Cache recovery is missing Biology, Integrated Math I, Integrated Math II, and Integrated Math III due to low matched sample sizes.

### Examine Distributions of Predicted Student Scores

The prediction equations for the non-disrupted samples provide a statistical statement about what to expect for students testing under non-disrupted conditions. The prediction is not perfect but, given the relatively high R<sup>2</sup> values, we can use the prediction equations derived in the non-disrupted samples to calculate how disrupted students might have scored had they not been disrupted. For each disrupted student, we computed their 2018 predicted score using the regression equation computed for the matched, non-disrupted students. Next, we computed the difference between the predicted scores and observed scores for the disrupted students, where positive values indicate higher predicted scores than observed and negative values indicate higher observed scores than predicted. These tables are presented in Appendix B. Table B1 presents the distribution of observed and predicted scores and the differences for the non-disrupted sample and Table B2 presents the differences between observed and predicted scores using the non-disrupted sample's equation for the multiple sign-in disruption. Tables B3 and B4 present the distributions for the over four hours disruption and Tables B5 and B6 present the distributions for the cache recovery disruption.<sup>5</sup>

### Compare Predictions of Disrupted Students to Non-Disrupted Students

*Large numbers* of students with *notable differences* between observed and predicted scores provides another piece of evidence about the impact of the computer disruptions. We defined large number and notable differences by comparing the difference in observed and predicted scores between the non-disrupted and disrupted samples. The non-disrupted sample represented the baseline: what would be expected under normal testing conditions.

First, we compared the distribution of differences using P-P plots. The P-P plots provide an evaluation of whether the differences between observed and predicted scores are normally distributed. Specifically, they plot the expected and observed cumulative distributions. We would

<sup>5</sup> Technical Note: When a prediction equation is derived on one sample and applied to a second sample, the variance of the residuals is expected to be larger due to shrinkage. Our predictions of performance for the disruption group were slightly weaker than would be expected based on the shrinkage associated with applying the prediction equation to a randomly equivalent sample. Given that our second sample was not randomly equivalent, but differs by the computer disruption, the small difference suggests that our prediction utility is not severely reduced in the disrupted sample.

expect the differences between observed and predicted scores to be normally distributed for the non-disrupted sample. That is, most of the differences should be near zero and there should be approximately equal numbers of differences where the observed score is greater than the predicted score and the predicted score is greater than the observed score. If the disruption impacted student test performance, then the difference between predicted and observed would be larger for the disrupted sample and deviate from both the normal distribution and the disrupted sample distribution. We compared P-P plots for the non-disrupted and disrupted samples. Appendix C provides the P-P plots. See Figure C9 for an example of little deviation of difference scores from the normal distribution and Figure C8 as an example of some deviation. Generally, the differences between predicted and observed scores varied from the normal distribution. This is indicated by the deviation of the tails of the distribution from 0 and suggests that, for both samples, there were some students that performed better or worse than expected. This effect may be due to the non-normal distribution of observed scale scores for some EOC exams. Most importantly, the plots are similar between the non-disrupted and disrupted samples. As such, there were no systematic differences between the distributions of the two samples.

Next, we computed the difference in observed and predicted scores at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> percentile for the non-disrupted sample and determined the percentage of students in the disrupted sample who were at or below the same cut point for the 5<sup>th</sup> and 10<sup>th</sup> non-disrupted percentile and those that were at or above the cut point for the 90<sup>th</sup> and 95<sup>th</sup> non-disrupted percentile. We performed this analysis by disruption type. If more than 5% and 10% of the disrupted students were below the 5<sup>th</sup> and 10<sup>th</sup> non-disrupted percentile cuts, respectively, then more students in the disrupted sample scored higher than expected. If more than 10% and 5% of the disrupted students were above the 90<sup>th</sup> and 95<sup>th</sup> non-disrupted percentile cuts, respectively, then more students in the disrupted sample scored lower than expected. Either case would provide evidence that the computer disruption had an impact on scores. Tables 9 through 11 present the percent of students in the disrupted sample below the 5<sup>th</sup> and 10<sup>th</sup> percentile cuts and above the 90<sup>th</sup> and 95<sup>th</sup> percentile cuts for the multiple sign-in, over four hours, and cache recovery disruptions, respectively.

For the multiple sign-in disruption, several EOC exams had a higher percent of disrupted students above the 90<sup>th</sup> and 95<sup>th</sup> cuts than would be expected and a lower number of students below the 5<sup>th</sup> and 10<sup>th</sup> percentile cuts than would be expected. The largest difference was observed for Integrated Math II, where 11.44% of disrupted students had higher predicted scores than observed scores; 6.44% higher than would have been expected based in the non-disrupted sample. It is important to note that, given our sample sizes for this EOC exam, these discrepancies amount to unexpected differences between predicted and observed scores for 24 - 25 students. A similar but smaller effect was observed for cache recovery and the over four hours disruptions. Overall, disrupted students had higher predicted scores than observed scores, providing evidence that the disruption had some negative impact.

**Table 9. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup> Percentile of Non-Disrupted Students for EOC Exams – Multiple Sign-In Disruption**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
Algebra I	2,695	4.16%	8.87%	12.57%	5.97%
Algebra II	1,350	3.70%	8.14%	11.70%	4.96%
Biology	3,567	4.51%	8.49%	11.21%	6.11%
Chemistry	3,210	4.61%	8.19%	12.31%	7.01%
English I	5,754	4.62%	9.68%	12.91%	6.95%
English II	4,879	4.71%	9.16%	12.75%	6.95%
English III	3,544	4.80%	9.15%	13.23%	8.38%
Geometry	2,104	3.71%	8.32%	11.55%	5.09%
Integrated Math I	506	2.96%	4.74%	12.05%	7.11%
Integrated Math II	376	6.65%	10.11%	17.56%	11.44%
Integrated Math III	274	5.11%	11.68%	15.69%	7.66%
U.S. History	2,030	4.14%	9.31%	12.51%	6.60%

Note. Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

**Table 10. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> Percentile of Non-Disrupted Students for EOC Exams – Over Four Hours Disruption**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
Algebra I	3,757	5.38%	10.25%	10.83%	5.16%
Algebra II	1,940	5.98%	11.24%	8.61%	4.64%
Biology	856	5.61%	11.68%	8.06%	4.67%
Chemistry	705	4.68%	7.23%	7.38%	4.40%
English I	10,473	4.89%	10.40%	11.00%	5.81%
English II	8,054	4.69%	9.99%	10.90%	5.87%
English III	5,352	5.16%	10.32%	12.22%	6.65%
Geometry	3,522	4.60%	10.25%	11.10%	4.94%
Integrated Math I	854	5.85%	10.42%	13.70%	7.38%
Integrated Math II	647	6.80%	11.75%	14.22%	8.35%
Integrated Math III	216	7.87%	12.04%	18.06%	9.26%
U.S. History	3,594	6.07%	11.13%	9.79%	5.45%

Note. Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

**Table 11. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> Percentile of Non-Disrupted Students for EOC Exams – Cache Recovery Disruption**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
Algebra I	544	2.57%	7.35%	11.58%	6.62%
Algebra II	681	5.29%	9.99%	8.96%	5.14%
Chemistry	712	4.49%	7.86%	11.80%	6.46%
English I	741	4.99%	9.04%	11.34%	7.29%
English II	201	3.48%	6.47%	15.92%	6.47%
English III	1,027	3.99%	9.25%	11.49%	6.43%
Geometry	206	3.88%	7.76%	14.56%	1.94%
U.S. History	969	1.34%	4.13%	11.25%	1.34%

*Note.* Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

### Student-Level Summary

Several analyses were conducted to examine the potential impact of three computer disruptions on student-level scores by EOC exam. The statistical evidence provided in this report is intended to inform the TNDOE about whether computer disruptions systematically impacted student test scores. The evidence presented thus far suggests that students who experienced the multiple sign-in disruption scored lower, on average, than students in the non-disrupted sample. Because we ruled out many other possible explanations for the difference by using propensity score matching, it is highly likely that the difference is due to the computer disruptions that occurred during the 2018 testing window. However, the analyses also suggest that the impact was not large, with score differences ranging from less than one point to six points. There was not a systematic effect for the over four hour and cache recovery disruptions. Although we observed lower scores for the over four hours and cache recovery disrupted samples on several of the EOC exams, the effect was not consistently detrimental, but at times were beneficial, with the disrupted sample scoring higher than the non-disrupted sample. Additionally, for some EOC exams, the differences were less than one score point, generally suggesting no difference between the two samples.

The level of the impact varied among the EOC exams. Across analyses, students experiencing multiple sign-in attempts scored lower than expected on every EOC exam. The largest differences were observed for Integrated Math II and English III exams. For these exams, the student-level means were lower, disruption helped predict 2018 EOC scores even after controlling for other variables, and disrupted students earned lower scores than what would have been expected given the non-disruption prediction model. For the over four hours disruption, the results were more mixed across EOC exams and evidence for a disruption effect was smaller. Integrated Math I, II, and III showed small detrimental effects, however, these results were based on low sample sizes, making conclusions regarding possible effects tenuous. U.S. History and Algebra II showed small beneficial effects, with the disrupted sample scoring one to three points higher than the non-disrupted sample. For the cache recovery disruption, the largest difference was observed for U.S. History. The disrupted sample scored, on average, three points lower than the non-disrupted sample and the disruption explained an additional 1.6% of the variance in test scores.

## School-Level Analyses

The TN DOE asked HumRRO to investigate whether the impact of disruptions on student scores, once aggregated, impacted school-level accountability results. We investigated the impact of computer disruption on school scores by examining alternative ways to compute school-level scores and proficiency, including and excluding students who experienced a disruption.

### School-Level Means

To evaluate the impact of disruption on school-level mean scores, we considered ways school-level means could be calculated including and excluding students who experienced a disruption. We used all available student-level data, including records that were removed from the student-level analyses because of missing data.

First, we computed school-level mean scores for each exam including all students from a school, and then computed the mean and standard deviation of the school-level means. Second, we removed from the “All Students” sample those students who were identified as having *any* computer disruption of any type. That is, the “No Disruptions” mean, included only students for which there was no evidence of a computer disruption. If a school had fewer than 10 students in either the “All Students” or the “No Disruptions” group, the school was removed from the mean of school means calculation. For some schools, the entire group of EOC students was considered disrupted resulting in no available students to compute the “No Disruptions” mean. These schools were also removed from the “All Students” mean. We removed these schools to provide more stable estimates for school means and ensure the comparison sample was based on the same set of schools. The sample sizes ( $n$ ) in Table 12 are the number of schools included in the mean of school means calculation. We compared “No Disruptions” to “All Students” so that a positive difference would indicate an *increase* in the mean after removing students who were disrupted.

The results in Table 12 suggest that, the mean *school-level* scores are slightly higher when students with disruptions are removed from the sample, as indicated by the positive Cohen’s  $d$ . This indicates that, in general, removing students who experienced computer disruptions from the school-level scores would result in a small positive effect.

**Table 12. School-Level Mean EOC Exam Scores by Group**

2018 EOC Exam	<i>n</i>	All Students Mean (SD)	No Disruptions Mean (SD)	Cohen's <i>d</i>
Algebra I	177	304.23 (24.28)	304.58 (24.82)	.01
Algebra II	141	292.55 (15.69)	293.06 (16.17)	.03
Biology	162	692.02 (23.47)	692.76 (23.17)	.03
Chemistry	162	675.35 (45.62)	677.17 (45.50)	.04
English I	144	321.70 (6.66)	322.11 (7.10)	.06
English II	149	309.73 (6.53)	310.28 (7.05)	.08
English III	150	321.52 (7.54)	322.07 (7.56)	.07
Geometry	148	299.97 (20.84)	300.46 (21.26)	.02
Integrated Math I	84	304.81 (26.95)	305.46 (27.13)	.02
Integrated Math II	56	298.63 (17.77)	298.96 (17.91)	.02
Integrated Math III	41	286.19 (20.71)	287.07 (20.91)	.04
U.S. History	157	328.04 (9.93)	328.32 (10.27)	.03

*Note.* *n* = Number of schools in the data with at least 10 students in both the “All Students” and “No Disruption” samples. The number of students within each school ranged from 41 to 177 depending on EOC exam.

### School-Level Classification

Next, we examined the effect of including and excluding students who experienced computer disruptions from the computation of school-level percent of students identified as being at least “On track” or proficient. Each EOC exam has four performance levels, and students are placed in one of them for each exam. We defined proficiency as scores in the upper two of the four performance levels for each EOC exam. For 10 of the 12 exams, these levels are labeled “On track” or “Mastered.” For the Biology and Chemistry exams, these levels are labeled “Proficient” and “Advanced.”

We began with the same data used for Table 12 which removed schools with fewer than 10 students in either the “All Students” or “No Disruptions” groups. We calculated the proportion of students defined as proficient for each school for each EOC exam, first across all students at each school, and then for only those with no evidence of a computer disruption. Then, we calculated the mean proportion among schools. The results are shown in Table 13.

Across all EOC exams, the percent of students who were at least “On track” was slightly higher when students who experienced a disruption were excluded from the school-level mean. The difference between the All Students and the “No Disruptions” students is generally very small, ranging from < 0.1 to 1.8 percentage points depending on the exam.

**Table 13. School-Level Percent Proficient Mean Scores by Group for EOC Exams**

2018 EOC Exam	<i>n</i>	All Students Mean (SD)	No Disruptions Mean (SD)	Difference
Algebra I	177	28.4% (28.6%)	28.7% (29.3%)	0.3%
Algebra II	141	19.6% (16.3%)	20.2% (17.0%)	0.7%
Biology	162	41.9% (21.1%)	42.3% (21.0%)	0.5%
Chemistry	162	30.1% (21.0%)	30.8% (21.4%)	0.6%
English I	144	21.4% (15.0%)	22.4% (16.3%)	1.0%
English II	149	27.7% (17.8%)	29.5% (19.0%)	1.8%
English III	150	23.5% (15.9%)	24.0% (16.0%)	0.5%
Geometry	148	26.0% (23.2%)	26.6% (23.6%)	0.6%
Integrated Math I	84	31.8% (29.5%)	32.5% (29.6%)	0.6%
Integrated Math II	56	26.5% (26.2%)	26.9% (26.5%)	0.4%
Integrated Math III	41	14.4% (18.8%)	14.5% (19.1%)	<0.1%
U.S. History	157	23.6% (17.6%)	24.0% (18.1%)	0.4%

Note. *n* = Number of schools in the data in both the “All Students” and “No Disruption” samples. The number of students within each school ranged from 41 to 177 depending on EOC exam.

### School-Level Summary

School-level accountability was based on the aggregation of student-level scores. Our investigation examined the effect of removing students that experienced a computer disruption from school-level means. Overall, our results indicated that excluding students who experienced a computer disruption resulted in higher school-level scores and more students classified as being at least “On track,” or proficient, for most EOC exams. For both comparisons, the differences were small, with school-level score differences ranging from .28 to 1.82 and classification differences ranging from < .1% - 1.8%.

### Invalidation

If a testing session was determined to be aberrant or irregular by the test administrator, he/she had discretion to invalidate the entire test administration and essentially delete the student scores from the official record. As a result, students whose scores were invalidated do not have 2018 EOC exam scores<sup>6</sup>. Test administrators did not, however, invalidate the scores for every student who experienced a computer disruption and had discretion in determining which students experienced irregular administrations. As a result of the computer disruption, the TN DOE expected a higher number of records to be invalidated in 2018 than in previous years and were concerned that a high volume of invalidated student records could impact aggregate-level (school, district, and state) results.

<sup>6</sup> It is important to note that the analyses described in earlier sections of this report, by definition, did not include any data from students whose scores were invalidated.



For these analyses, the TNDOE was interested in answering the following questions:

- To what extent did invalidation of student records impact data?
- Are invalidated testing records representative of the student population at a school, district, and state-level? Prior year testing population?
- Are invalidated testing records associated with student characteristics?

To address the first question, we examined the percent of students whose scores were invalidated across EOC exams. Generally, invalidation was rare, with between .06% and 4.06% of records identified as irregular. At most, 532 of 12,581 students (4.06%) taking the Integrated Math II exam had invalidated scores. Because the rate of invalidation was small, we do not expect that the invalidation impacted 2018 scores. However, we are unable to address that question directly because we cannot isolate the impact of invalidation on 2018. For the 2018 scores, any invalidation effects are confounded with the known disruption events that occurred in 2018. Additionally, changes have occurred in the TNReady administration over time (e.g., changes in administration requirements, participation rates, existing legislation). Specifically, in April of 2018, following the system-wide occurrence of disruptions and prior to the end of the TNReady testing window, the Tennessee General Assembly passed House Bill No. 1981 and House Bill No. 75, effectively eliminating negative impacts of 2018 TNReady exams on students, teachers, and schools. The introduction of this legislation may have lowered students' motivation to perform well.

To address the TNDOE's concerns over whether invalidated test records were representative of the student population and associated with student characteristics, we examined the distribution of gender, race, and indicators of student economic disadvantage, special education, and ELL status by invalidation status. We created five 2 x k tables that compared the number of students with validated records (0 = validated, 1 = invalidated) to the membership in the demographic variable, where k is the number of levels in the demographic variable. For example, invalidation and gender have two levels each so the table was 2 (validated or invalidated) X 2 (male or female). A chi-square ( $\chi^2$ ) test of independence compares the distribution of two or more categorical variables to determine if there is a statistically significant difference in observed frequencies and expected frequencies, where the expected frequencies are equivalent proportions of student characteristics for the validated and invalidated group. Chi-square tests are affected by large sample sizes and can result in inflated Type I error rates. Therefore, due to the large sample sizes, we also computed the phi coefficient ( $\phi$ ) as an effect size. Phi can be interpreted like a correlation between two categorical variables. In other words, does gender correlate with invalidation status in the 2 x 2 table mentioned previously? The results of this analysis for gender, race, economic disadvantage, special education, and ELL status are in Table 14.

Gender appears to be unrelated to invalidation status as indicated by the non-significant chi-square and very small phi coefficients. The other four demographic variables have some statistically significant results for some EOC exams, but all phi coefficients are small ( $< .06$ ) suggesting that the distribution of these demographics was the same for the invalidated and validated groups. The largest effects, although still practically small, was for race and economically disadvantaged status.

**Table 14. Gender and Race by Irregular Administration Comparisons for EOC Exams**

Content	Gender		Race		Economically Disadvantaged		Special Education		ELL Status	
	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$
Algebra I	3.47	-.01	9.10	.01	5.14*	.01	1.62	.01	0.01	.00
Algebra II	0.15	.00	6.94	.01	1.97	-.01	22.08***	.02	48.59***	.03
Biology	0.92	.00	22.32***	.02	9.38**	-.01	0.65	.00	9.58**	.01
Chemistry	1.62	.01	44.86***	.03	0.00	.00	4.67*	-.01	0.39	.00
English I	0.03	.00	45.72***	.03	99.26***	-.04	11.16***	-.01	9.59**	-.01
English II	0.04	.00	52.10***	.03	88.59***	-.04	9.51**	-.01	3.89*	-.01
English III	1.41	-.01	97.57***	.05	87.75***	-.04	18.33***	-.02	2.44	-.01
Geometry	0.02	.00	18.24**	.02	20.33***	-.02	0.06	.00	1.65	.01
Int. Math I	0.50	-.01	46.57***	.06	52.62***	-.06	5.25*	-.02	29.37***	-.05
Int. Math II	0.71	.01	22.75***	.04	43.55***	-.06	6.13*	-.02	5.22*	-.02
Int. Math III	0.97	.01	2.67	.02	1.97	.02	0.06	.00	0.07	.00
U.S. History	0.08	.00	69.32***	.04	7.61**	.01	2.76	.01	2.55	.01

Note. Int. Math = Integrated Math.  $\chi^2$  = Chi-Square,  $\phi$  = Phi coefficient. ELL = English Language Learner.  
 \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

We were unable to address several parts of the second invalidation question. Specifically, because the rate of invalidation across the state was small, the school and district-level invalidation, on average, was also small, which prohibited our ability to make reasonable inferences as to whether the students invalidated at the school or district-level were representative of the school or district population. We were also unable to answer the question about whether invalidated testing records were representative of prior year testing population. To answer this question, we need student records for those that were invalidated in 2017. The instructions for test administrators to indicate an “irregular administration” were different in 2017 than in 2018, so there is no corresponding indicator in the 2017 to draw a comparison.

### **Invalidation Summary**

Overall, the rate of invalidation was rare, impacting 6,838 of 499,735 EOC student records across the state. The demographic characteristics of students whose scores were invalidated for 2018 were very similar to the other students who tested in 2018, with no differences on gender, special education and ELL status and very small differences on race and economically disadvantaged status. This suggests that invalidation records were generally representative of the state population and were not associated with specific student characteristics.

## Grade-Level Results

Computer disruptions also occurred for some students who tested online during the TNReady grades 5 through 8 exams. The TNDOE requested that HumRRO examine the effects of the three disruption types on grade-level exam scores. We conducted the same analyses as we did with EOC exams with some slight differences in methodology noted below.

### *Data Cleaning*

The TN DOE provided 370,852 grades 5 through 8 student records for four 2018 grade-level exams that were administered on the computer: ELA, Math, Science, and Social Studies. We only examined grades 5 through 8 because grade 3 and 4 exams were not administered on the computer. Similar to the EOC exams, the TN DOE provided scale scores, demographic variables, and indicators for each of the disruption types<sup>7</sup> and had already performed some data screening, for example, removing home-schooled students from the data. After removing records with missing scale scores and demographic variables, we retained 358,325 student records.

We used 2017 ELA, Math, and Science scores as the prior-year test score for the grade-level exams. Social Studies test scores were not provided for 2017, as this was a field test year. With grade-level exams, most students take the same exam as they progress through grades, allowing us to match 2018 ELA, Math, and Science scores with 2017 ELA, Math, and Science scores, respectively. That is, the prior-year test for grade 5 Math is grade 4 Math. Because there were no data for Social Studies 2017 test scores, we used 2017 ELA test scores for each grade. This prior-year test was chosen as the closest in theoretical content (i.e., reading ability). Table 15 lists the 2017 exam scores that were matched with each 2018 grade-level exam. Table 15 also contains the total student sample size (Total  $n$  with scores), the student sample size with complete data (2018 & 2017 Merged  $n$ ), and the student sample size experiencing each of the three computer disruption types.

As with EOC, some students experienced more than one disruption type and were included in each sample. The analyses were conducted separately for each type of disruption and each of grade-level exam. The multiple sign-in disruption was defined in the same way as the EOC exams—if a student was disrupted on at least two subparts of the exam, the student was considered disrupted. Students who were only disrupted on one subpart were set aside for these analyses. The inclusion of students in the over four hour and cache recovery disruptions was the same as the EOC exams. The number of grade 5 students experiencing a cache recovery disruption was too small ( $n < 100$ ) for stable statistical comparisons.

---

<sup>7</sup> TN DOE request the primary test vendor identify students that experienced the three disruption types.

**Table 15. Merged Total Sample Sizes and Disruption Type Sample Sizes by 2018 Grade-Level Exam**

2018 Grade-Level Exam	Total <i>n</i> with scores	Most Relevant 2017 Exam	2018 & 2017 Merged <i>n</i>	Multiple Sign-in <sup>a</sup>		Over Four Hours		Cache Recovery <sup>b</sup>	
				<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
ELA Gr. 5	11,831	ELA Gr. 4	11,221	1,303	11.61%	2,761	24.61%	--	--
ELA Gr. 6	25,891	ELA Gr. 5	24,468	2,849	11.64%	7,361	30.08%	108	0.44%
ELA Gr. 7	26,585	ELA Gr. 6	25,133	2,293	9.12%	6,428	25.58%	127	0.51%
ELA Gr. 8	26,528	ELA Gr. 7	25,041	2,006	8.01%	5,531	22.09%	113	0.45%
MAT Gr. 5	11,828	MAT Gr. 4	11,233	645	5.74%	331	2.95%	--	--
MAT Gr. 6	25,436	MAT Gr. 5	24,016	1,332	5.55%	1,289	5.37%	134	0.56%
MAT Gr. 7	26,553	MAT Gr. 6	25,170	1,274	5.06%	1,292	5.13%	167	0.66%
MAT Gr. 8	23,529	MAT Gr. 7	22,206	1,002	4.51%	908	4.09%	122	0.55%
SCI Gr. 5	11,814	SCI Gr. 4	11,224	218	1.94%	522	4.65%	--	--
SCI Gr. 6	25,688	SCI Gr. 5	24,263	395	1.63%	920	3.79%	132	0.54%
SCI Gr. 7	26,372	SCI Gr. 6	24,989	366	1.46%	941	3.77%	177	0.71%
SCI Gr. 8	25,990	SCI Gr. 7	24,619	390	1.58%	939	3.81%	138	0.56%
SOC Gr. 5	11,812	ELA Gr. 4	11,205	316	2.82%	693	6.18%	--	--
SOC Gr. 6	25,648	ELA Gr. 5	24,229	793	3.27%	1,450	5.98%	132	0.54%
SOC Gr. 7	26,195	ELA Gr. 6	24,748	705	2.85%	979	3.96%	174	0.70%
SOC Gr. 8	26,625	ELA Gr. 7	25,130	575	2.29%	1,120	4.46%	135	0.54%

Note. ELA = English Language Arts. MAT = Math. SCI = Science. SOC = Social Studies. Gr. = Grade.

<sup>a</sup>Disruption was defined as experiencing at least one multiple sign-in attempt on more than one subpart of the exam.

<sup>b</sup>Excluded exams for which the sample size was too small ( $n < 100$ ).

### Propensity Score Matching

Propensity score matching was used to match disrupted grade-level students with their non-disrupted peers that were similar on several variables. The propensity score model was identical to the one used in the EOC exams except for the grade variable. We did not match on grade-level because all but a few students testing for the grade-level exam were the same grade. The model consisted of the following individual and school-level variables:

- Race
- Sex
- Relevant prior test scores
- English Language Learner status
- Special Education status
- Economically disadvantage status
- School-level achievement
- School-level proportion of economically disadvantage students

The propensity score matching process was identical to the EOC exam process. Each disrupted student was matched to a non-disrupted student who was similar on the variables listed above. Prior to matching, the largest differences across the grades and subjects by disruption was Science grade 8 Special Education status ( $d = .49$ ), Social Studies grade 8 school-level prior-year test mean score ( $d = .67$ ), and ELA grade 8 school-level prior-year test mean score ( $d = -.44$ ) for multiple sign-in, over four hours, and cache recovery, respectively.

After matching, we examined mean differences between the disrupted and matched non-disrupted samples on the matching variables. The mean effect size (Cohen's  $d$ ) across all matching variables was .00 (range =  $-.11$  to  $.12$ ), .00 (range =  $-.29$  to  $.14$ ), and .01 (range =  $-.21$  to  $.24$ ) for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. The largest differences after matching was for Science grade 5 economic disadvantage school-level proportion for over four hours ( $d = -.29$ ) and ELA grade 8 economic disadvantage student-level indicator for cache recovery ( $d = .24$ ). Although these differences surpassed  $|.20|$  for Cohen's  $d$ , the differences in proportion amounted to 3% and 9% for these two variables, respectively, between disrupted and non-disrupted matched samples. We did not anticipate these minor differences to occlude the results of the analyses using these matched samples. Covariate comparisons before and after matching are found in Appendix D for grade-level exams<sup>8</sup>.

### Student-Level Analyses

Using the matched samples for each disruption type, we examined whether test scores of students who experienced computer disruptions differed from the test scores of students who were not disrupted. By matching the samples on several variables known to be related to student exam scores, we controlled for the impact of these matching variables on group differences. As with EOC exams, we first examined the mean differences by disruption type and grade-level exam.

#### Mean Exam Score Comparisons

Mean comparisons of the 2018 grade-level scale scores are presented in Tables 16, 17, and 18 for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. Means and SD were compared using an independent-samples  $t$ -test ( $t$ -value) and Cohen's  $d$  effect size estimates. For the multiple sign-in disruption, the mean score for the disrupted sample was lower than the mean score for the matched non-disrupted sample for all 2018 grade-level exams. Across grades, the effect was largest for Science and smallest for Social Studies. The largest effect size difference was observed for Science grade 8 with a mean score difference of 10.5 scale score points and a Cohen's  $d$  value of  $-.23$ .

For the over four hours disruption, the mean score differences were also small with Cohen's  $d$  effect sizes ranging from  $-.10$  to  $.18$ . Across subjects, Science exams had the largest negative mean differences, but the grade 5 exam showed disrupted students performing better than their matched non-disrupted peers. Science exams had the lowest sample sizes which may make conclusions regarding possible effects tenuous for this subject. As with EOC exams, many of the differences were less than one score point for the over four hours disruption, suggesting no difference between the two samples.

---

<sup>8</sup> We used Cohen's  $d$  to compare proportions of dichotomous variables (i.e., gender, race) to assess for balance between the two samples. We also provide  $t$ -tests to compare all demographic variables and prior year test performance.

For the cache recovery disruption, all effect sizes were small, ranging from -.24 to .01. However, by and large, the results showed the disrupted sample scoring lower than the non-disrupted sample for all but one exam. Although the EOC results for the cache recovery disruption was mixed, for grade-level exams, there appears to be a consistent negative difference in mean scale scores between the disrupted and matching non-disrupted samples. The largest difference was for Science grade 8 with a mean difference of 10.3 scale score point and a Cohen's *d* of -.24.

**Table 16. Mean Grade-Level Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Multiple Sign-In Disruption**

2018 Grade-Level Exam	Disrupted		Non-Disrupted		<i>t</i> -value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
ELA Gr. 5	1,303	316.24 (34.45)	1,303	316.87 (33.02)	-0.48	-.02
ELA Gr. 6	2,849	325.38 (29.25)	2,849	329.28 (29.03)	-5.05***	-.13
ELA Gr. 7	2,293	323.46 (31.79)	2,293	324.73 (29.86)	-1.40	-.04
ELA Gr. 8	2,006	320.18 (31.34)	2,006	322.35 (31.58)	-2.19*	-.07
Math Gr. 5	645	313.54 (49.39)	645	320.05 (51.40)	-2.32*	-.13
Math Gr. 6	1,332	318.05 (36.21)	1,332	323.64 (36.77)	-3.95***	-.15
Math Gr. 7	1,274	316.71 (42.37)	1,274	319.54 (40.11)	-1.73	-.07
Math Gr. 8	1,002	303.78 (42.27)	1,002	308.63 (40.64)	-2.62**	-.12
Science Gr. 5	218	744.46 (45.55)	218	750.79 (49.29)	-1.39	-.13
Science Gr. 6	395	751.71 (46.46)	395	757.28 (45.34)	-1.71	-.12
Science Gr. 7	366	749.22 (45.22)	366	755.59 (44.22)	-1.93	-.14
Science Gr. 8	390	739.15 (43.11)	390	749.67 (46.48)	-3.28**	-.23
Social Studies Gr. 5	316	315.23 (29.69)	316	317.74 (27.53)	-1.10	-.09
Social Studies Gr. 6	793	319.03 (31.82)	793	321.57 (27.92)	-1.69	-.08
Social Studies Gr. 7	705	311.45 (32.47)	705	313.86 (28.15)	-1.49	-.08
Social Studies Gr. 8	575	324.62 (31.68)	575	327.17 (27.52)	-1.46	-.09

*Note.* ELA = English Language Arts. Gr. = Grade. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on more than one subpart of the exam.

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

**Table 17. Mean Grade-Level Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Over Four Hours Disruption**

2018 EOC Exam	Disrupted		Non-Disrupted		<i>t</i> -value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
ELA Gr. 5	2,761	321.15 (31.25)	2,761	321.89 (31.49)	-0.87	-.02
ELA Gr. 6	7,361	335.57 (28.86)	7,361	336.81 (27.87)	-2.65**	-.04
ELA Gr. 7	6,428	335.79 (29.96)	6,428	334.65 (28.93)	2.19*	.04
ELA Gr. 8	5,531	334.88 (30.68)	5,531	331.62 (29.78)	5.67***	.11
Math Gr. 5	331	322.72 (43.59)	331	326.68 (50.31)	-1.08	-.08
Math Gr. 6	1,289	325.77 (35.11)	1,289	327.05 (35.75)	-0.92	-.04
Math Gr. 7	1,292	323.92 (41.24)	1,292	323.08 (39.47)	0.53	.02
Math Gr. 8	908	314.09 (42.86)	908	312.80 (41.55)	0.65	.03
Science Gr. 5	522	768.96 (45.62)	522	761.05 (44.02)	2.85**	.18
Science Gr. 6	920	771.78 (43.86)	920	774.82 (43.06)	-1.50	-.07
Science Gr. 7	941	767.41 (41.91)	941	771.36 (42.74)	-2.02*	-.09
Science Gr. 8	939	755.31 (40.07)	939	759.48 (43.13)	-2.17*	-.10
Social Studies Gr. 5	693	323.11 (29.83)	693	319.79 (27.14)	2.17*	.12
Social Studies Gr. 6	1,450	327.95 (29.79)	1,450	326.84 (27.34)	1.04	.04
Social Studies Gr. 7	979	322.85 (30.67)	979	320.46 (27.39)	1.82	.08
Social Studies Gr. 8	1,120	339.14 (31.97)	1,120	334.25 (28.16)	3.83***	.16

Note. ELA = English Language Arts. Gr. = Grade. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on one subpart of the exam.

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

**Table 18. Mean Grade-Level Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Cache Recovery Disruption**

2018 EOC Exam	Disrupted		Non-Disrupted		<i>t</i> -value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
ELA Gr. 5 <sup>a</sup>	--	--	--	--	--	--
ELA Gr. 6	108	328.98 (25.44)	108	332.86 (26.74)	-1.09	-.15
ELA Gr. 7	127	327.39 (29.28)	127	329.97 (27.85)	-0.72	-.09
ELA Gr. 8	113	329.51 (28.69)	113	330.56 (26.49)	-0.28	-.04
Math Gr. 5 <sup>a</sup>	--	--	--	--	--	--
Math Gr. 6	134	327.19 (37.49)	134	331.02 (38.10)	-0.83	-.10
Math Gr. 7	167	320.43 (41.47)	167	320.15 (38.68)	0.06	.01
Math Gr. 8	122	317.03 (45.42)	122	320.07 (38.52)	-0.57	-.07

(continued)

**Table 18. Mean Grade-Level Exam Scores of Disrupted Students and Matched Non-Disrupted Students for Cache Recovery Disruption (continued)**

2018 EOC Exam	Disrupted		Non-Disrupted		<i>t</i> -value	Cohen's <i>d</i>
	<i>n</i>	Mean (SD)	<i>n</i>	Mean (SD)		
Science Gr. 5 <sup>a</sup>	--	--	--	--	--	--
Science Gr. 6	132	758.50 (41.01)	132	761.80 (39.16)	-0.67	-.08
Science Gr. 7	177	750.93 (44.30)	177	761.27 (41.88)	-2.26*	-.24
Science Gr. 8	138	754.03 (44.49)	138	759.36 (38.70)	-1.06	-.13
Social Studies Gr. 5 <sup>a</sup>	--	--	--	--	--	--
Social Studies Gr. 6	132	318.24 (27.76)	132	320.05 (32.70)	-0.49	-.06
Social Studies Gr. 7	174	308.01 (31.19)	174	311.67 (27.09)	-1.17	-.13
Social Studies Gr. 8	135	324.18 (25.31)	135	327.96 (26.16)	-1.21	-.15

Note. ELA = English Language Arts. Gr. = Grade. *t*-value = Test of mean difference. Negative Cohen's *d* values indicate a lower mean score for the disrupted sample. A student was considered disrupted if he or she experienced a disruption on one subpart of the exam.

<sup>a</sup>Group sample sizes were too small (*n* < 100) to present a stable mean for comparison.

\**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

### 2018 Exam Score Predictability

Next, we examined if disruption added to the predictability of the 2018 test scores beyond other known indicators of performance. The covariates included in this model were prior year test scores, sex, race, prior year school-level test scores, school-level economically disadvantaged status, and student-level indicators of English Language Learner, economically disadvantaged, or special education status. If a computer disruption impacted scores, then inclusion of the disruption indicator (0 = not disrupted, 1 = disrupted) in the multiple regression model should add to the estimation of 2018 scores, as indicated by the size of the multiple regression coefficient ( $R^2$ ). The addition of any variable to a regression model should increase the multiple regression coefficient ( $R^2$ ) but it is the degree of increase that indicates if that variable contributes to the prediction of the test score. We conducted regression modeling analyses separately for each grade-level exam and each disruption type where sample size permitted.

Table 19 shows the  $R^2$  values for each regression model with and without the disruption variable, the change in  $R^2$  between the two models ( $\Delta R^2$ ), and the unstandardized regression coefficient for disruption (*b*).  $R^2$  values can be interpreted as the proportion of test score variance explained by the model. An  $R^2$  of .621 indicates 62.1% of the variance in ELA grade 6 test scores is explained by all variables mentioned above. After adding the group membership variable for disruption (0 = not disrupted, 1 = disrupted) to the model, the variance explained increases to .624 or a .003 change ( $\Delta R^2$ ). This  $\Delta R^2$  reflects a .3% increase in variance explained, which is small. The unstandardized coefficient for disruption (*b*) is on the metric of the grade-level scale scores and indicates the scale score difference between those not disrupted and those disrupted, after controlling for the other variables in the model. Many of these values were statistically significant, as noted by the test of the regression coefficient (*t*-value). However, these coefficients can be statistically significant due to large sample sizes. Therefore, overall effects should be interpreted considering the range of each grade-level exam scores and the standardized effect size (in this case,  $\Delta R^2$ ) which, at most, accounted for an additional 1.3% of the variance in 2018 grade-level scores. The range of grade-level scale scores is 200-450 for all exams except for Science, which ranges from 600-900.



As with EOC exams, for all disruption types, there was very little or no change in the R<sup>2</sup> values when the disruption variable was included in the regression model, indicating that disruption added very little to the prediction of 2018 grade-level scores beyond the other factors known to impact test scores. The direction of the unstandardized regression coefficients (*b*) was generally negative meaning that disruption lowered observed scores, controlling for the other variables in the model. The regression coefficients were largest in magnitude for the multiple sign-in disruption. This result coincides with the individual-level mean differences.

**Table 19. Incremental Validity Estimation of Disruption by Disruption Type and Grade-Level Exam**

Disruption Type	Grade-Level Exam	<i>n</i>	Covariates		$\Delta R^2$	Disruption <i>b</i> estimate	<i>t</i> -value
			Covariates Only R <sup>2</sup>	+ Disruption R <sup>2</sup>			
Multiple Sign-In	ELA Gr. 5	2,606	.625	.625	.000	-0.743	-0.91
	ELA Gr. 6	5,698	.621	.624	.003	-2.940	-6.18***
	ELA Gr. 7	4,586	.659	.659	.000	-1.121	-2.10*
	ELA Gr. 8	4,012	.654	.655	.001	-1.858	-3.16**
	MAT Gr. 5	1,290	.711	.715	.004	-6.227	-4.13***
	MAT Gr. 6	2,664	.670	.677	.008	-6.496	-8.04***
	MAT Gr. 7	2,548	.688	.689	.001	-2.432	-2.66**
	MAT Gr. 8	2,004	.610	.612	.003	-4.191	-3.61***
	SCI Gr. 5	436	.700	.701	.002	-3.869	-1.53
	SCI Gr. 6	790	.626	.630	.004	-5.833	-2.89**
	SCI Gr. 7	732	.653	.659	.006	-7.088	-3.61***
	SCI Gr. 8	780	.674	.687	.013	-10.143	-5.55***
	SOC Gr. 5	632	.555	.555	.000	-0.570	-0.37
	SOC Gr. 6	1,586	.525	.528	.003	-3.276	-3.15**
	SOC Gr. 7	1,410	.554	.556	.002	-3.023	-2.79**
	SOC Gr. 8	1,150	.561	.562	.002	-2.333	-2.00*
Over Four Hours	ELA Gr. 5	5,522	.613	.613	.000	-0.026	-0.05
	ELA Gr. 6	14,722	.619	.620	.001	-1.822	-6.29***
	ELA Gr. 7	12,856	.665	.665	.000	0.405	1.35
	ELA Gr. 8	11,062	.656	.657	.001	2.198	6.51***
	MAT Gr. 5	662	.710	.711	.001	-2.971	-1.49
	MAT Gr. 6	2,578	.682	.684	.002	-3.055	-3.88***
	MAT Gr. 7	2,584	.673	.674	.001	2.107	2.32*
	MAT Gr. 8	1,816	.644	.645	.000	1.257	1.06
	SCI Gr. 5	1,044	.678	.679	.002	3.773	2.35*
	SCI Gr. 6	1,840	.633	.633	.000	-1.327	-1.08
	SCI Gr. 7	1,882	.695	.696	.001	-3.130	-2.89**
	SCI Gr. 8	1,878	.630	.632	.002	-3.717	-3.17**

(continued)

**Table 19. Incremental Validity Estimation of Disruption by Disruption Type and Grade-Level Exam (continued)**

Disruption Type	Grade-Level Exam	<i>n</i>	Covariates		$\Delta R^2$	Disruption <i>b</i> estimate	<i>t</i> -value
			Covariates Only $R^2$	+ Disruption $R^2$			
Over Four Hours (cont'd)	SOC Gr. 5	1,386	.526	.528	.002	2.827	2.67**
	SOC Gr. 6	2,900	.504	.504	.000	0.844	1.12
	SOC Gr. 7	1,958	.573	.573	.000	1.195	1.38
	SOC Gr. 8	2,240	.562	.567	.005	4.291	5.08***
Cache Recovery	ELA Gr. 6	216	.586	.594	.008	-4.635	-1.98*
	ELA Gr. 7	254	.637	.637	.000	0.180	0.08
	ELA Gr. 8	226	.658	.659	.001	1.514	0.68
	MAT Gr. 6	268	.752	.757	.005	-5.105	-2.18*
	MAT Gr. 7	334	.698	.699	.001	1.870	0.75
	MAT Gr. 8	244	.673	.673	.000	1.684	0.53
	SCI Gr. 6	264	.560	.561	.001	-1.880	-0.56
	SCI Gr. 7	354	.685	.696	.012	-9.354	-3.60***
	SCI Gr. 8	276	.683	.688	.006	-6.299	-2.18*
	SOC Gr. 6	264	.513	.513	.000	-0.681	-0.26
SOC Gr. 7	348	.542	.543	.002	-2.301	-1.06	
SOC Gr. 8	270	.546	.551	.004	-3.509	-1.60	

Note. ELA = English Language Arts. MAT = Math. SCI = Science. SOC = Social Studies. Gr. = Grade.  
*t*-value = Test of the unstandardized regression coefficient for Disruption *b* estimate. Results for Grade 5 ELA, Math, Science, or Social Studies exams are not presented for the cache recovery disruption due to low sample sizes in the disrupted sample. Covariates in the model were prior-year test scores, sex, race, indicators for students who are English Language Learners, Economically Disadvantaged, or Special Education, and school-level prior-year test means and school-level Economically Disadvantaged status.  
 \**p* < .05. \*\**p* < .01. \*\*\**p* < .001.

Next, as with EOC results, we estimated regression equations to predict test scores *separately* for each sample (disrupted and non-disrupted) for each grade-level exam and disruption type. Recall, if students' performance was affected by the computer disruption, the strength of the prediction, as indicated by  $R^2$ , should be lower for the disrupted students than for the non-disrupted students. A lower  $R^2$  coefficient means that students' performance in the disrupted sample was not predicted as well as the non-disrupted sample. This is yet another way to gauge whether there was a general impact across students due to the computer disruptions.

Tables 20 through 22 present the  $R^2$  values for the disrupted and non-disrupted samples for the multiple sign-in, over four hours, and cache recovery disruptions, respectively. Overall, 2018 test scores were fairly well predicted for both disrupted and non-disrupted samples, with 47% to 76% of the variance accounted for by the available set of predictor variables. These  $R^2$  values are higher for grade-level exams compared to EOC, which is likely due to the higher congruency of the prior-year test scores in the model. Generally, there were slightly higher  $R^2$  values for the non-disrupted samples, for all three disruption types, as shown by the positive  $R^2$  difference values. This means that grade-level exam scores were not quite as well predicted by the available set of prediction variables (or covariates), and suggests the computer disruption did, perhaps, have some impact on exam scores. The difference in variance accounted for ranged from .1% to 13.4% across all comparisons, suggesting that the impact was small, except for Social Studies grade 6, cache recovery which was 20.4%. This difference could be anomalous

due to small sample size or it could suggest that, the cache recovery disruption did have a relatively larger impact on our ability to predict scores for this exam. For some grade-level exams, the  $R^2$  values were higher for the disrupted sample compared to the non-disrupted sample, with differences ranging from  $< .1\%$  to  $10.8\%$ .

As with EOC exams, we computed RMSE estimates to account for differences in sample variance. Recall, RMSE is the root of the variance of the residuals between observed and predicted scores and shows us how well the model “fits” or replicates the observed scores. The RMSE differences were small (compared to the RMSE values), ranging from  $-3.4$  to  $.91$ . This indicates that the model fit for the two samples was similar and for some grade-level exams the model replicated the observed score better for the non-disrupted sample and for other grade-level exams the model replicated the observed scores better for the disrupted sample.

**Table 20. Predictability of 2018 Grade-Level Scale Scores for Non-Disrupted and Disrupted Groups for Multiple Sign-In Disruption**

Content	<i>n</i>	Non-Disrupted $R^2$	Disrupted $R^2$	$R^2$ Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
ELA Gr. 5	1,303	.628	.627	.001	20.23	21.13	-0.89
ELA Gr. 6	2,849	.626	.620	.005	17.80	18.06	-0.26
ELA Gr. 7	2,293	.647	.671	-.023	17.78	18.29	-0.51
ELA Gr. 8	2,006	.661	.651	.010	18.45	18.57	-0.13
Math Gr. 5	645	.721	.713	.008	27.38	26.69	0.70
Math Gr. 6	1,332	.696	.659	.038	20.36	21.25	-0.89
Math Gr. 7	1,274	.699	.685	.014	22.12	23.90	-1.78
Math Gr. 8	1,002	.622	.607	.015	25.14	26.65	-1.52
Science Gr. 5	218	.744	.662	.082	25.52	27.12	-1.60
Science Gr. 6	395	.635	.631	.004	27.74	28.57	-0.83
Science Gr. 7	366	.649	.686	-.037	26.56	25.76	0.80
Science Gr. 8	390	.699	.672	.027	25.87	25.05	0.82
Social Studies Gr. 5	316	.531	.586	-.055	19.16	19.45	-0.29
Social Studies Gr. 6	793	.474	.582	-.108	20.41	20.73	-0.33
Social Studies Gr. 7	705	.567	.557	.010	18.68	21.79	-3.10
Social Studies Gr. 8	575	.573	.562	.011	18.16	21.18	-3.02

Note. ELA = English Language Arts. Gr. = Grade.  $R^2$  Difference is the difference between the Non-Disrupted  $R^2$  and the Disrupted  $R^2$ .

**Table 21. Predictability of 2018 Grade-Level Scale Scores for Non-Disrupted and Disrupted Groups for Over Four Hours Disruption Type**

Content	<i>n</i>	Non-Disrupted R <sup>2</sup>	Disrupted R <sup>2</sup>	R <sup>2</sup> Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
ELA Gr. 5	2,761	.612	.617	-.005	19.67	19.39	0.28
ELA Gr. 6	7,361	.609	.632	-.023	17.45	17.53	-0.08
ELA Gr. 7	6,428	.651	.679	-.027	17.10	17.00	0.10
ELA Gr. 8	5,531	.655	.658	-.003	17.51	17.96	-0.45
Math Gr. 5	331	.754	.676	.078	25.32	25.23	0.10
Math Gr. 6	1,289	.681	.691	-.010	20.29	19.62	0.67
Math Gr. 7	1,292	.661	.689	-.028	23.10	23.12	-0.02
Math Gr. 8	908	.650	.651	-.001	24.76	25.48	-0.73
Science Gr. 5	522	.686	.679	.008	24.94	26.16	-1.22
Science Gr. 6	920	.647	.627	.020	25.75	26.97	-1.22
Science Gr. 7	941	.695	.705	-.010	23.77	22.90	0.86
Science Gr. 8	939	.645	.617	.028	25.86	24.95	0.91
Social Studies Gr. 5	693	.519	.555	-.035	18.98	20.08	-1.10
Social Studies Gr. 6	1,450	.497	.515	-.018	19.47	20.84	-1.36
Social Studies Gr. 7	979	.569	.583	-.014	18.10	19.92	-1.83
Social Studies Gr. 8	1,120	.564	.573	-.009	18.68	20.99	-2.31

Note. ELA = English Language Arts. Gr. = Grade. R<sup>2</sup> Difference is the difference between the Non-Disrupted R<sup>2</sup> and the Disrupted R<sup>2</sup>.

**Table 22. Predictability of 2018 Grade-Level Scale Scores for Non-Disrupted and Disrupted Groups for Cache Recovery Disruption Type**

Grade-level Exam	<i>n</i>	Non-Disrupted R <sup>2</sup>	Disrupted R <sup>2</sup>	R <sup>2</sup> Difference	Non-Disrupted RMSE	Disrupted RMSE	RMSE Difference
ELA Gr. 6	108	.673	.539	.134	15.97	18.05	-2.07
ELA Gr. 7	127	.687	.623	.065	16.23	18.75	-2.52
ELA Gr. 8	113	.672	.706	-.033	15.73	16.22	-0.49
Math Gr. 6	134	.762	.769	-.007	19.34	18.74	0.60
Math Gr. 7	167	.704	.725	-.022	21.72	22.42	-0.70
Math Gr. 8	122	.632	.737	-.104	24.39	24.34	0.04
Science Gr. 6	132	.595	.573	.022	25.94	27.88	-1.94
Science Gr. 7	177	.728	.686	.041	22.50	25.54	-3.04
Science Gr. 8	138	.688	.722	-.035	22.47	24.45	-1.99
Social Studies Gr. 6	132	.615	.411	.204	21.11	22.16	-1.05
Social Studies Gr. 7	174	.567	.541	.026	18.37	21.77	-3.41
Social Studies Gr. 8	135	.561	.552	.009	18.10	17.69	0.42

Note. ELA = English Language Arts. Gr. = Grade. R<sup>2</sup> Difference is the difference between the Non-Disrupted R<sup>2</sup> and the Disrupted R<sup>2</sup>. Cache recovery is missing ELA, Math, Science, and Social Studies grade 5 exams due to low matched sample sizes.

## ***Examine Distributions of Predicted Student Scores***

As with EOC exams, we used the prediction equations for the non-disrupted samples to calculate how disrupted students might have scored had they not been disrupted. For each disrupted student, we computed their predicted grade-level exam score using the regression equation computed for the matched, non-disrupted students. Next, we computed the difference between the predicted scores and observed scores for the disrupted students, where positive values indicate higher predicted scores than observed and negative values indicate higher observed scores than predicted. These tables are presented in Appendix E. Table E1 presents the distribution of observed and predicted scores and the differences for the non-disrupted sample and Table E2 presents the differences between observed and predicted scores using the non-disrupted sample's equation for the multiple sign-in disruption. Tables E3 and E4 present the distributions for the over four hours disruption and Tables E5 and E6 present the distributions for the cache recovery disruption.

## ***Compare Predictions of Disrupted Students to Non-Disrupted Students***

First, we compared the distribution of differences using P-P plots. The P-P plots provide an evaluation of whether the differences between observed and predicted scores are normally distributed. Specifically, they plot the expected and observed cumulative distributions. We would expect the differences between observed and predicted scores to be normally distributed for the non-disrupted sample. That is, most of the differences should be near zero and there should be approximately equal numbers of differences where the observed score is greater than the predicted score and the predicted score is greater than the observed score. If the disruption impacted student test performance, then the difference between predicted and observed would be larger for the disrupted sample and deviate from both the normal distribution and the disrupted sample distribution. We compared P-P plots for the non-disrupted and disrupted samples. Appendix F provides the P-P plots for grade-level exams. Generally, the differences between predicted and observed scores varied from the normal distribution. This is indicated by the deviation of the tails of the distribution from 0 and suggests that, for both samples, there were some students that performed better or worse than expected. As with EOC exams, this effect may be due to the non-normal distribution of observed scale scores for some grade-level exams. Most importantly, the plots were similar between the non-disrupted and disrupted samples. As such, there were no systematic differences between the distribution of the two samples.

Recall, we considered the regression model for the matched, non-disrupted sample for each exam to represent the baseline: what would be expected under normal testing conditions. We computed the difference in observed and predicted scores at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> percentile for the non-disrupted sample and determined the percentage of students in the disrupted sample who were at or below the same cut point for the 5<sup>th</sup> and 10<sup>th</sup> non-disrupted percentile and those that were at or above the cut point for the 90<sup>th</sup> and 95<sup>th</sup> non-disrupted percentile. Essentially, this analysis compared differences between predicted and observed scores. If more than 5% and 10% of the disrupted students were below the 5<sup>th</sup> and 10<sup>th</sup> non-disrupted percentile cuts, respectively, then more students in the disrupted sample scored higher than expected. If more than 10% and 5% of the disrupted students were above the 90<sup>th</sup> and 95<sup>th</sup> non-disrupted percentile cuts, respectively, then more students in the disrupted sample scored lower than expected. Either case would provide evidence that the computer disruption had an impact on scores. Tables 23 through 25 present the percent of students in the disrupted sample below the 5<sup>th</sup> and 10<sup>th</sup> percentile cuts and above the 90<sup>th</sup> and 95<sup>th</sup> percentile cuts for the multiple sign-in, over four hours, and cache recovery disruptions, respectively.

For the multiple sign-in disruption, several grade-level exams had a higher percent of disrupted students above the 90<sup>th</sup> and 95<sup>th</sup> cuts than would be expected and a lower number of students below the 5<sup>th</sup> and 10<sup>th</sup> percentile cuts than would be expected. This effect was more extreme for the grade-level exams than the EOC exams. The largest difference was observed for Science grade 8, where 20.26% of disrupted students had higher predicted scores than observed scores; 10.26% higher than would have been expected based in the non-disrupted sample. Given our sample sizes, this equates to unexpected differences for 40 students. A similar but smaller effect was observed for cache recovery and the over four hours disruptions. As with EOC exams, disrupted students had higher predicted scores than observed scores overall, providing evidence that the disruption had some negative impact.

**Table 23. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup> Percentile of Non-Disrupted Students – Multiple Sign-In Disruption for Grade-Level Exams**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
ELA Gr. 5	1,303	4.99%	11.59%	10.82%	6.06%
ELA Gr. 6	2,849	3.26%	6.84%	13.33%	6.77%
ELA Gr. 7	2,293	4.14%	8.63%	12.78%	6.37%
ELA Gr. 8	2,006	3.99%	9.37%	11.26%	5.73%
Math Gr. 5	645	2.95%	7.45%	14.58%	5.74%
Math Gr. 6	1,332	2.10%	5.93%	17.79%	10.66%
Math Gr. 7	1,274	4.95%	8.72%	12.32%	6.51%
Math Gr. 8	1,002	4.49%	9.38%	13.08%	7.49%
Science Gr. 5	218	6.42%	11.47%	14.67%	9.17%
Science Gr. 6	395	3.54%	6.83%	15.70%	8.61%
Science Gr. 7	366	2.46%	4.92%	17.22%	9.84%
Science Gr. 8	390	2.05%	3.59%	20.26%	10.26%
Social Studies Gr. 5	316	4.11%	13.29%	12.34%	4.75%
Social Studies Gr. 6	793	6.05%	8.19%	13.24%	8.95%
Social Studies Gr. 7	705	5.25%	8.65%	16.60%	8.94%
Social Studies Gr. 8	575	6.26%	10.61%	16.70%	9.57%

*Note.* ELA = English Language Arts. Gr. = Grade. Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

**Table 24. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> Percentile of Non-Disrupted Students – Over Four Hours Disruption for Grade-Level Exams**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
ELA Gr. 5	2,761	5.40%	10.98%	9.20%	4.56%
ELA Gr. 6	7,361	4.17%	8.30%	11.31%	5.92%
ELA Gr. 7	6,428	4.57%	10.00%	9.93%	4.87%
ELA Gr. 8	5,531	6.53%	12.71%	8.78%	4.19%
Math Gr. 5	331	6.04%	10.57%	12.38%	7.85%
Math Gr. 6	1,289	2.87%	7.29%	14.12%	8.77%
Math Gr. 7	1,292	6.11%	12.38%	9.60%	3.95%
Math Gr. 8	908	7.16%	12.89%	9.36%	5.95%
Science Gr. 5	522	7.66%	12.83%	8.81%	4.60%
Science Gr. 6	920	4.67%	9.56%	11.08%	5.43%
Science Gr. 7	941	2.76%	7.75%	9.99%	6.06%
Science Gr. 8	939	3.73%	7.56%	11.72%	6.71%
Social Studies Gr. 5	693	7.94%	14.72%	9.81%	3.61%
Social Studies Gr. 6	1,450	6.00%	10.00%	9.45%	4.76%
Social Studies Gr. 7	979	7.35%	12.36%	11.65%	6.64%
Social Studies Gr. 8	1,120	9.20%	15.81%	8.22%	5.54%

Note. ELA = English Language Arts. Gr. = Grade. Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

**Table 25. Percent of Disrupted Students with Predicted and Observed Scale Score Differences at the 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup> and 95<sup>th</sup> Percentile of Non-Disrupted Students – Cache Recovery Disruption for Grade-Level Exams**

Content	<i>n</i>	5 <sup>th</sup>	10 <sup>th</sup>	90 <sup>th</sup>	95 <sup>th</sup>
ELA Gr. 6	108	1.85%	7.41%	20.37%	14.81%
ELA Gr. 7	127	11.02%	16.53%	16.53%	7.87%
ELA Gr. 8	113	7.08%	13.27%	15.92%	9.73%
Math Gr. 6	134	5.97%	11.19%	16.42%	9.70%
Math Gr. 7	167	10.18%	14.97%	10.78%	6.59%
Math Gr. 8	122	7.38%	14.76%	9.02%	6.56%
Science Gr. 6	132	8.33%	14.39%	15.91%	6.82%
Science Gr. 7	177	2.26%	8.47%	10.17%	2.26%
Science Gr. 8	138	3.62%	6.52%	20.29%	13.04%
Social Studies Gr. 6	132	9.09%	12.12%	12.88%	6.06%
Social Studies Gr. 7	174	2.87%	7.47%	16.67%	8.62%
Social Studies Gr. 8	135	1.48%	7.41%	17.04%	8.15%

Note. ELA = English Language Arts. Gr. = Grade. Percentages larger than 5% or 10% at the 5<sup>th</sup> and 10<sup>th</sup> percentile, respectively, indicates that more disrupted students than expected earned a higher observed than predicted score. Percentages larger than 5% and 10% at the 95<sup>th</sup> and 90<sup>th</sup> percentile, respectively, indicates that more disrupted students earned a lower observed score than expected.

## Student-Level Summary

Several analyses were conducted to examine the potential impact of three computer disruptions on student-level scores by grade-level exam. The evidence presented thus far for grade-level exams suggested that students who experienced the multiple sign-in disruption scored lower, on average, than students in the non-disrupted sample. This effect was larger than the effect observed on the EOC exams. However, similar to the EOC exam results, there was not a systematic effect for the over four hour and cache recovery disruptions.

The level of the impact varied among the grade-level exam and disruption type. Across analyses, students experiencing multiple sign-in attempts scored lower than expected on every grade-level exam. For these exams, the student-level means were lower, disruption helped predict 2018 grade-level scores even after controlling for other variables, and disrupted students earned lower scores than what would have been expected given the non-disruption prediction model. This effect was larger for grade-level exams compared to EOC exams. Disrupted students scored 4.1 scale score points lower, on average across all grade-level exams, whereas disrupted students on the EOC exams scored 2.6 points lower compared to their non-disrupted peers. The largest grade-level difference was for grade 8 Science exam where disrupted students scored 10.5 points lower, on average, than their matched non-disrupted peers. For the over four hours disruption, the results were more mixed across grade-level exams and evidence for a disruption effect was smaller. Disruption in general was rarer for grade-level exams making conclusions regarding possible effects tenuous for some comparisons. We encourage the TN DOE to consider the sample sizes for each analysis, with larger sample sizes indicating more stable effects.

## School-Level Analyses

The TN DOE asked HumRRO to investigate whether the impact of disruptions on student scores, once aggregated, impacted school-level accountability results. We investigated the impact of computer disruption on school scores by examining alternative ways to compute school-level scores and proficiency, including and excluding students who experienced a disruption.

### School-Level Means

We examined the impact of disruptions on school-level mean scores by including and excluding students who experienced a disruption. We used all available student-level data, including records that were removed from the student-level analyses because of missing data. These analyses follow the same methodology used for the EOC exams. First, we computed school-level means for each exam including all students from a school, and then computed the mean and standard deviation of the school-level means. Second, we removed from the “All Students” sample those students who were identified as having *any* computer disruption of any type. That is, the “No Disruptions” mean included only students for which there was no evidence of a computer disruption. If a school had fewer than 10 students in either the “All Students” or the “No Disruptions” sample, the school was removed from the mean of school means calculation. For some schools, the entire group of students was considered disrupted resulting in no available students to compute the “No Disruptions” mean. These schools were also removed from the “All Students” mean. We removed these schools to provide more stable estimates for school means and ensure the comparison sample was based on the same set of schools. The sample sizes ( $n$ ) in Table 26 are the number of schools included in the mean of school means



calculation. We compared “No Disruptions” to “All Students” so that a positive difference would indicate an *increase* in the mean after removing students who were disrupted.

The results in Table 26 suggest that, the mean *school-level* scores are slightly higher when students with disruptions are removed from the sample, as indicated by the positive Cohen’s *d*. This indicates that, in general, removing students who experienced computer disruptions from the school-level scores would result in a small positive effect. These effect sizes are larger, on average, for the grade-level exams compared to the EOC exams.

**Table 26. School-Level Mean Grade-Level Exam Scores by Group**

2018 Grade-Level Exam	<i>n</i>	All Students Mean (SD)	No Disruptions Mean (SD)	Cohen’s <i>d</i>
ELA Grade 5	55	324.77 (14.68)	325.61 (14.62)	.06
ELA Grade 6	69	332.68 (9.63)	334.55 (9.65)	.19
ELA Grade 7	71	330.57 (10.46)	330.85 (10.10)	.03
ELA Grade 8	73	327.77 (11.59)	327.84 (12.25)	.01
Math Grade 5	57	336.77 (24.45)	338.71 (24.18)	.08
Math Grade 6	74	329.60 (13.74)	331.47 (13.94)	.13
Math Grade 7	74	323.52 (16.10)	325.08 (15.80)	.10
Math Grade 8	77	316.50 (17.46)	318.54 (17.99)	.11
Science Grade 5	57	763.60 (19.80)	764.60 (19.68)	.05
Science Grade 6	73	766.35 (16.79)	767.99 (16.26)	.10
Science Grade 7	72	765.24 (16.71)	767.03 (16.38)	.11
Science Grade 8	75	759.41 (17.95)	760.63 (18.06)	.07
Social Studies Grade 5	57	325.84 (15.80)	326.73 (15.55)	.06
Social Studies Grade 6	72	322.41 (11.01)	322.90 (11.52)	.04
Social Studies Grade 7	73	313.88 (10.19)	314.28 (10.14)	.04
Social Studies Grade 8	75	326.20 (11.27)	326.50 (11.03)	.03

*Note.* ELA = English Language Arts. *n* = Number of schools in the data with at least 10 students in both the “All Students” and “No Disruption” samples.

### School-Level Classification

Next, we examined the effect of including and excluding students who experienced computer disruptions from the computation of school-level percent of students identified as being at least “On track” or proficient. As with the EOC exams, each grade-level exam has four performance levels, and students are placed in one of them for each exam. We defined proficiency as scores in the upper two of the four performance levels for each grade-level exam. For ELA and Math, these levels are labeled “On track” or “Mastered.” For the Science exams, these levels are labeled “Proficient” and “Advanced.” Social studies did not have proficiency levels assigned as the standard setting for these exams was scheduled for the summer of 2018.

We began with the same data used for Table 26 which removed schools with fewer than 10 students in either the “All Students” or “No Disruptions” samples. We calculated the proportion of students defined as proficient for each school and grade-level exam, first across all students at each school, and then for only those with no evidence of a computer disruption. Then, we calculated the mean proportion among schools. The results are shown in Table 27.

Across most grade-level exams, the percent of students who were at least “On track” was slightly higher when students who experienced a disruption were excluded from the school-level mean. The exception was three ELA grades which had < 1% decrease in percent proficient, a difference of essentially zero. The difference between the All Students and the No Disruptions group is generally small, ranging from 0.7% to 2.3% depending on the exam. This difference is somewhat larger compared to the EOC exam results.

**Table 27. School-Level Percent Proficient Mean Scores by Group for Grade-Level Exams**

2018 Grade-Level Exam	<i>n</i>	All Students Mean (SD)	No Disruptions Mean (SD)	Difference
ELA Grade 5	55	39.0% (18.8%)	38.9% (19.4%)	-0.2%
ELA Grade 6	69	39.2% (14.1%)	41.5% (14.8%)	2.3%
ELA Grade 7	71	37.1% (15.2%)	36.5% (14.8%)	-0.7%
ELA Grade 8	73	28.6% (13.8%)	28.3% (15.2%)	-0.3%
Math Grade 5	57	50.4% (23.1%)	52.0% (22.9%)	1.6%
Math Grade 6	74	41.7% (16.2%)	43.8% (16.5%)	2.1%
Math Grade 7	74	34.8% (17.6%)	35.9% (18.0%)	1.1%
Math Grade 8	77	39.4% (17.7%)	41.5% (18.3%)	2.0%
Science Grade 5	57	64.7% (18.4%)	65.4% (18.2%)	0.7%
Science Grade 6	73	65.2% (16.2%)	67.2% (15.5%)	2.0%
Science Grade 7	72	65.6% (15.5%)	67.3% (15.2%)	1.6%
Science Grade 8	75	62.2% (17.9%)	63.4% (18.0%)	1.2%

Note. ELA = English Language Arts. *n* = Number of schools in the data in both the “All Students” and “No Disruption” samples. Social Studies did not have performance levels assigned to scale scores.

### School-Level Summary

School-level accountability was based on the aggregation of student-level scores. Our investigation examined the effect of removing students that experienced a computer disruption from school-level means. Overall, our results indicated that excluding students who experienced a computer disruption resulted in higher school-level scores and more students classified as being at least “On track,” or proficient, for most grade-level exams. For both comparisons, the differences were small, with school-level score differences ranging from < .10 to 2.03 and classification differences ranging from -0.7% to 2.3%.

### Invalidation

If a testing session was determined to be aberrant or irregular by the test administrator, he/she had discretion to invalidate the entire test administration and essentially delete the student scores from the official record. As with EOC exams, the TN DOE expected a higher number of grade-level exam records to be invalidated in 2018 than in previous years and were concerned that a high volume of invalidated student records could impact aggregate-level (school, district, and state) results.

We examined the percent of students whose scores were invalidated across grade-level exams. Generally, invalidation was rarer than for the EOC exams, with < 1% of records identified as irregular across all grades. Several grade-level exams had sample sizes of invalidated students < 100. Grade-level exams with less than 100 students are not presented as the results could be potentially misleading. To address the TNDOE’s concerns over whether invalidated test records

were associated with student characteristics and representative of the state-level population, we examined the distribution of gender, race, and indicators of student economic disadvantage, special education, and ELL status by invalidation status. As we did with EOC exams, we created five  $2 \times k$  tables that compared the number of students with validated records (0 = validated, 1 = invalidated) to the membership in the demographic variable, where  $k$  is the number of levels in the demographic variable. Recall, invalidation and gender have two levels each, so the table was  $2$  (validated or invalidated)  $\times$   $2$  (male or female). A chi-square ( $\chi^2$ ) test of independence compares the distribution of two or more categorical variables to determine if there is a statistically significant difference in observed frequencies and expected frequencies, where the expected frequencies are equivalent proportions of student characteristics for the validated and invalidated sample. Chi-square tests are affected by large sample sizes and can result in inflated Type I error rates. Therefore, due to the large sample sizes, we also computed the phi coefficient ( $\phi$ ) as an effect size. Phi can be interpreted like a correlation between two categorical variables. In other words, does gender correlate with invalidation status in the  $2 \times 2$  table mentioned previously? The results of this analysis for gender, race, economic disadvantage, special education, and ELL status are in Table 28. Several grade-level exams were excluded from these analyses because invalidation status was very rare ( $n < 50$ ), which could result in sparseness, or too few individuals in each cell of the  $2 \times k$  tables.

Gender appeared to be unrelated to invalidation status as indicated by the non-significant chi-square and very small phi coefficients. The other four demographic variables have some statistically significant results for some EOC exams, but all phi coefficients are small ( $< .04$ ) except for one, suggesting that the distribution of these demographics was the same for the invalidated and validated samples. The largest effect was for race for the Math grade 8 exam, but this effect was still practically small ( $\phi = .11$ ).

**Table 28. Gender and Race by Irregular Administration Comparisons for Grade-Level Exams**

Content	Gender		Race		Economically Disadvantaged		Special Education		ELL Status	
	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$	$\chi^2$	$\phi$
ELA Grade 6	1.81	-.01	38.81***	.04	0.00	.00	1.41	.01	1.41	.01
ELA Grade 7	0.70	.01	4.97	.01	2.04	-.01	9.23**	.02	9.23**	.02
ELA Grade 8	2.37	.01	12.94*	.02	0.06	.00	0.84	.01	0.84	.01
Math Grade 8	0.02	.00	293.42***	.11	17.14***	.03	0.71	-.01	0.71	-.01
Social Studies Gr. 7	0.31	.00	6.32	.02	0.87	-.01	7.03**	-.02	7.03**	-.02
Social Studies Gr. 8	0.68	.00	14.01*	.02	1.70	-.01	3.55	-.01	3.55	-.01

*Note.* ELA = English Language Arts. Gr. = Grade.  $\chi^2$  = Chi-Square,  $\phi$  = Phi coefficient. ELL = English Language Learner.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

### Invalidation Summary

Invalidation was rare for grade-level exams in 2018 with less than 1% of students with invalidated scores (978 of 277,224 student records). The demographic characteristics of students whose scores were invalidated for 2018 were very similar to the other students who tested in 2018, with no differences on gender, race, special education, ELL status and economically disadvantaged status. A very small difference in race for Math grade 8 was observed but this difference was practically very small ( $\phi = .11$ ) to warrant further evaluation.

This suggests that invalidation records were generally representative of the state population and were not associated with specific student characteristics.

## General Conclusions

Several events occurred that disrupted the testing experiences for some students who tested online during the 2018 TNReady testing window. Specifically, three types of disruptions occurred for at least some students who tested online: (a) students were involuntarily signed out during the test session and had to re-initiate the sign-in process at least once, (b) students lost connectivity or were booted off of the system and signed in again at a later time or date to finish their exam, resulting in over four hours elapsing between initial sign-in and test submission, and (c) due to system errors, a student's data was lost during the test session and the test administrator had to request recovery from the computer cache. Because of the wide-spread and systematic nature of these computer disruptions, the TNDOE wished to examine the impact of these disruptions on student performance. This report is intended to inform the TNDOE of the statistical impact of computer disruptions and invalidation on 2018 EOC and grade-level exam scores. These findings can be used to inform decisions about whether any policy actions are appropriate.

We examined each disruption type separately and by EOC and grade-level exam. We noted evidence for small, directional effects in several EOC and grade-level exams where disrupted students earned lower scores than their non-disrupted matched peers and performed worse than expected. These effects were more pronounced across grade-level exams, with students across grades 5 through 8 scoring lower than their matched non-disrupted peers. Because we ruled out many other possible explanations for the difference by using propensity score matching, it is highly likely that the difference is due to the computer disruptions that occurred during the 2018 testing windows. This effect was particularly noteworthy for the multiple sign-in disruption, where students scored lower than expected on every EOC and grade-level exam. This effect was also observed in the school-level aggregate means.

There was not a systematic effect for the over four hour and cache recovery disruptions, but cache recovery was somewhat more detrimental to students taking the grade-level exams. Although we observed lower scores for the disrupted sample on several TNReady exams for the over four hours and cache recovery disruptions, the effect was not consistently detrimental, but at times the disruption seemed beneficial, with the disrupted sample scoring higher than the non-disrupted sample. Additionally, for some EOC and grade-level exams, the differences were less than one scale score point, generally suggesting no meaningful difference between the two samples.

Overall, systematic effects were observed where students who experienced multiple sign-in attempts tended to earn lower scores than their matched peers who were not disrupted. Scores for students who experienced this disruption should be interpreted with these results in mind.

## References

- Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulation. *Biometrical Journal*, *51*(1), 171-184. doi: 10.1002/bimj.200810488
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155
- Connelly, B. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*, *66*, 407-442. doi: 10.1111/peps.12020