Final Report on Online Interruptions of the Spring 2015 Smarter Balanced Assessment Administration in Montana, Nevada, and North Dakota

September 2, 2016

Joseph Martineau
Nathan Dadey

Center for Assessment

EXECUTIVE SUMMARY

Montana, North Dakota, and Nevada students took the Smarter Balanced assessment in spring of 2015. All three states experienced interruptions in the online assessment. While these three states reported that students experienced multiple types of interruptions during testing, the data systems did not, apparently, capture data necessary to study many types of interruptions. These issues with data limited the analyses to one type of interruption – students being logged out, involuntarily, due to a server failure. Data issues also precluded more sensitive examinations of the effects of interruptions as a function of when a student was logged out during the test (i.e., at what point in the test the interruption occurred, which is a measure of the severity of an interruption). Therefore, this report only addresses involuntary logout interruptions without regard to the severity of the interruption. The analysis, thus, only captures a portion of the total interruptions students experienced.

A relatively small percentage of students experienced logout interruptions - no more than 6.2 percent of students experienced a logout interruption in any combination of subject, grade and state. The effects of logout interruptions on student scores were generally small and ranged from strong negative effects (the strongest main effect of interruptions is a standardized mean difference of -0.43, after controlling for interactions with demographics) to strong positive effects (a covariate adjusted standardized mean difference of 0.31). The size and direction of the effects, in general, displayed little consistent pattern across states, grades, subject areas, and demographic groups. The effects in lower grades and in mathematics were often negative, but again, this pattern is weak.

Clearer patterns emerge for when looking at participation. Compared to baseline rates in the previous three school years, the percentage of enrolled students who received valid test scores tended to be appreciably lower than in the past, at both the school and district level, and those results were consistent across subjects, grades, and states. Across subjects, grades and states the percentages of students who received valid scores during the 2014-2015 year decreased by 5 to 21 percent at the school-level and 6 to 16 percent at the district level, relative to the baseline percentage from the prior three years. In addition, these decreases were significant in a large percentage of the schools and districts. Across subjects, grades and states the percentage of schools with a significant decrease ranged from 16 percent to 36 percent and the percentage of districts ranged from 17 to 38 percent.

To address the limitations encountered during the analysis, this report also provides recommendations on ways to improve systems that administer, and record the results of, online assessments.

INTRODUCTION

This report provides a summary of the investigation conducted by The National Center for the Improvement of Educational Assessment (Center for Assessment) staff regarding the interruptions to online testing that occurred during the spring 2015 administration of the Smarter Balanced assessments in Montana, Nevada, and North Dakota. The first section provides background information. The second section details issues regarding the access to and existence of needed data elements to fully address interruptions. These data issues limited the breadth of analyses. The third section presents a revised analyses plan approach and the results of analyses designed to quantify the effect of interruptions. In the fourth and final section, recommendations are given for the design or redesign of online assessment platforms and related database systems to circumvent the difficulties encountered in this investigation.

BACKGROUND

Defining Interruptions

In the spring of 2015, three states administering the Smarter Balanced assessment in English language arts (ELA) and mathematics experienced both system-wide and idiosyncratic interruption events. We broadly define an interruption as "an event that disrupts students' testing experiences caused by the computers, online systems, or other technological devices through which the test is delivered" (Martineau, Domaleski, Egan, Patelis & Dadey, 2015, p. 5). These interruptions affect a student's ability to interact with the test administration platform in the manner intended and may be detrimental to the student's performance on the assessment. When taking a test, a student is interacting not only with a test administration platform on a specific device, such as a laptop or tablet, but also a number of other software applications, data systems, servers, and connectivity systems implemented by the vendor and state. Errors within this complex system of technology can occur, and when these events occur, students' testing experience may be interrupted.

We define systematic events as those that affected all students taking the test in a given state for a specific period of time. System-wide interruptions occurred in all three states (Montana, Nevada, and North Dakota), generally attributable to servers overloaded to the degree that they fail to respond to client device requests in a reasonably timely manner. During these system-wide events, all students logged onto the online testing platform were involuntarily logged off. Also, during these events no student was able to log onto the platform to start or resume testing. Idiosyncratic events were also reported in all three states. We define idiosyncratic interruptions as those occurring on a case-by-case basis outside the windows of time covered by system-wide interruptions. Based on conversations with state and vendor staff, we believe these types of interruptions occurred due to factors idiosyncratic to the student's specific testing experience, including slowed server responses (rather than server failure), selection of certain accessibility tools and the presentation of poorly rendered items. In terms of the latter two factors - accessibility tools and poorly rendered items - reports indicated that the test interface became unresponsive after their selection. Table 1 displays the potential types of causes of interruptions (across the top) and types of interruptions (down the left side). The check marks in the table represent the types of interruptions that could be caused by the types of causes.

The cells displayed in white show the types of causes and interruptions this project is intended to address (i.e., the scope of this project is limited to evaluating the effects of server failures and server overloads). The restriction of this work to the white cells dealing with system-wide interruptions stems from expected limitations in the available data.

While the analyses is restricted to system-wide slowdowns, in the remainder of this section we detail what data is needed to examine each type of interruption, then contrast with that data was available in the

following section. The complete set of types of data useful for analyzing the effect of any type of interruption is described below:

A. Yes/no interruption indicator for each student[1]
B. Interruption event start and end timestamps[2]
C. Measure of the severity of each interruption event[3]
D. Item-level timestamps for each student showing[4]
   - First navigation to (and last navigation away from) the page on which the item is displayed
   - First (and last) time the item was displayed on the student's screen
   - First (and last) time the student interacted with any part of the item

Table 1. *Relationship between potential causes and types of interruptions.*

	Type of Cause					
	Vendor Servers		Other Vendor Software/Hardware			External to Vendor Systems
Type of Interruption	Server Failure	Server Overload	Bug in Interface	Bug in Content	Other Bug In System	
Involuntary log out	✔	✔	✔	✔	✔	✔
Inability to log in	✔	✔	✔		✔	✔
Slow response & voluntary log out		✔	✔	✔	✔	✔
Slow response & continue working		✔	✔	✔	✔	✔
Frozen device & restart			✔	✔	✔	✔
Inaccurate content display			✔	✔	✔	✔
Inaccurate interaction w/interface			✔	✔	✔	✔
Scope of Interruption	**Systematic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**

Table 2 replicates Table 1, but in place of the checkmarks, indicates which of the types of data (A, B, C, and D) are useful for analysis of impact.

Table 2. *Types of data needed to analyze impact of interruptions.*

	Type of Cause					
	Vendor Servers		Other Vendor Software/Hardware			External to Vendor Systems
Type of Interruption	Server Failure	Server Overload	Bug in Interface	Bug in Content	Other Bug In System	
Involuntary log out	ABD	ABD	ABD	ABD	ABD	ABD
Inability to log in	ABD	ABD	ABD		ABD	ABD
Slow response & voluntary log out		ABCD	ABCD	ABCD	ABCD	ABCD
Slow response & continue working		ABCD	ABCD	ABCD	ABCD	ABCD
Frozen device & restart			ABD	ABD	ABD	ABD
Inaccurate content display			ABD	ABD	ABD	ABD
Inaccurate interaction w/interface			ABD	ABD	ABD	ABD
Scope of Interruption	**Systematic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**	**Idiosyncratic**

Figure 1 demonstrate the usefulness of the combination of data types B, i.e., event start and end timestamp data, and D, i.e., item time-stamp data, in analyzing the effect of interruptions for hypothetical two students.

---

[1] To indicate whether a student experienced an interruption at some point in the testing process.
[2] An ideal set of data to determine which item(s) were taken before, during, and/or after an interruption. Minimally acceptable data would be a timestamp for the start of an interruption.
[3] To allow for investigating the effect of different severity of interruptions (e.g., latency measures can represent how long a student had to wait between clicking on a button to proceed to the next item and the next item being displayed).
[4] An ideal set of data to determine which item(s) were taken before, during, and/or after an interruption. Minimally acceptable data would be any one of the following: first or last time the item was displayed on the student's screen, first or last time the student interacted with any part of the item.

In these figures, time is represented across the top (at the very left, the test has just begun, and at the very right is the last time the student was engaged with the test). The top rows (displayed in light blue) are used to represent times in which students experienced interruptions (the specific times that interruptions were experienced are displayed in gray). The top two rows displayed in light green describe and number the interruption events. The bottom two rows displayed in light green show the item number(s) associated with each item event, and describe the events. The bottom rows displayed in light blue indicate for each of the 10 items in the hypothetical test two periods of time:

- When the student was on the page where the item was located (the cells in light gray).
- The period between when the item was first displayed on the student's screen and the time when the student last interacted with any of the item's content (the cells in darker gray).

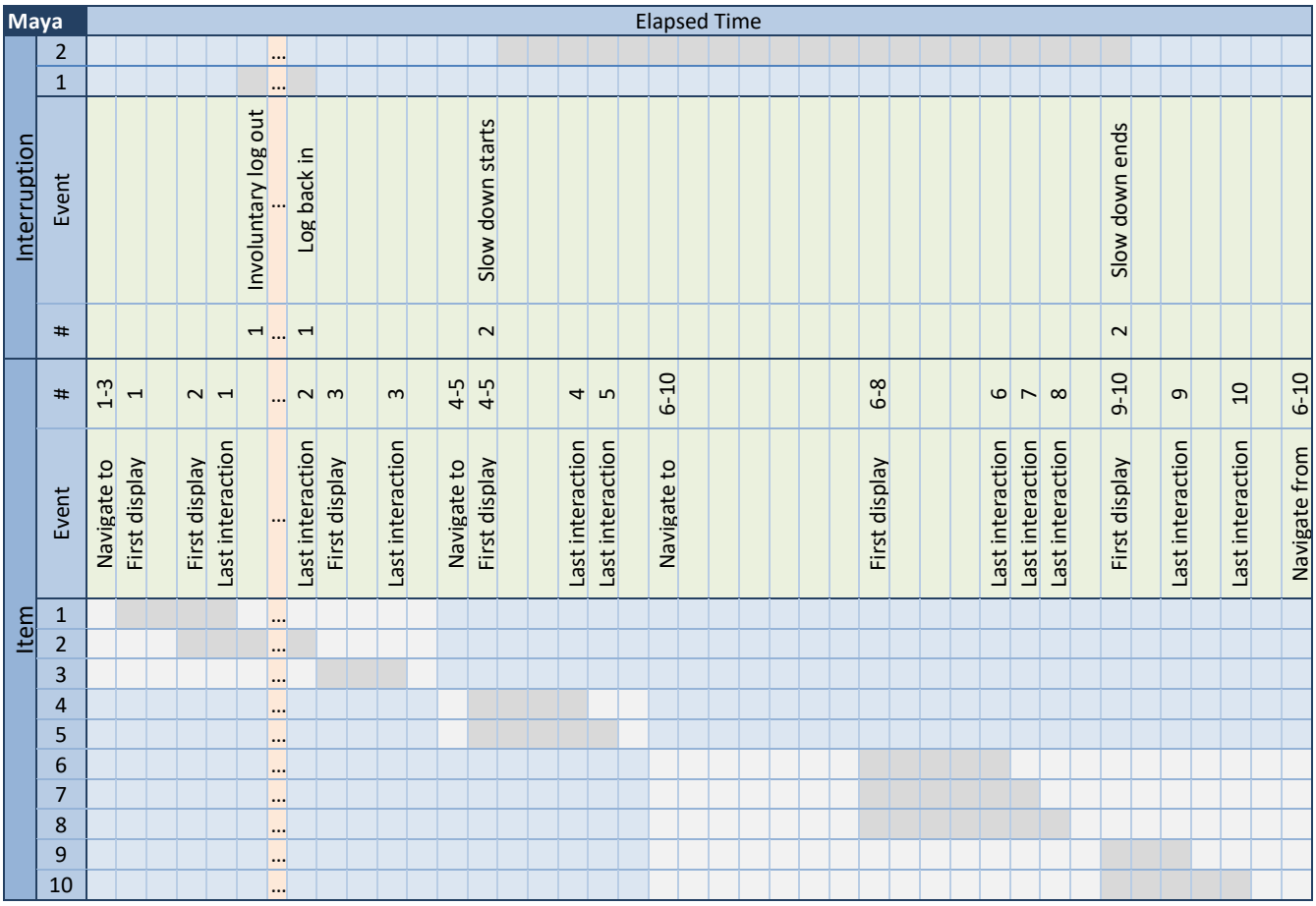

Figure 1. Sample timeline for a hypothetical student (Maya) in a 10-item test.

In Figure 1, it can be seen that Maya began her test normally and continued through his test normally through completing item 1 and beginning work on item 2. At that point, the system failed and logged him out involuntarily. She logged back in some time later and finished responding to question 2, and proceeded normally through item 5. A system slowdown affected the time required to load items 6-10. She continued working through the slowdown and completed the test.

Using the data about the times of interruptions (gray bars in the top rows) and the times of engaging with items (gray bars in the bottom rows) it is possible to tell which item may be affected by an interruption (e.g., any item with a gray bar that overlaps with or is after the gray bar representing the time of the interruption).

Interruptions can affect individual student scores, but they can also affect the rates at which students participate and complete testing. The effects on participation and completion can arise from decisions made by either local or state leaders regarding whether students must complete testing (if they have already started) or begin testing (if they have not already started). For example, one can imagine scenarios in which a student, who has unsuccessfully attempted to take a test multiple times, is not required to complete testing.

DATA LIMITATIONS

Unexpected Limitations in Available Data

Through conversations between staff at the Center for Assessment, States, Smarter Balanced (the supporter of and client for the development of the open-source testing platform), Measured Progress (the operational implementer of the open-source platform for test administration), and AIR (the developer of the open-source platform), it became clear that there are two salient limitations that prevent completion of intended analyses: the apparent (i) missing timestamps and (ii) inability to identify idiosyncratic interruptions. The following two sections discuss these apparently[5] missing data elements.

Timestamps

Timestamps indicating when each student first saw an item, when the student first selected a response to that item, when the student clicked to go onto the next item, or anything reasonably similar do not appear to be available. The available timestamps instead represent the time at which the item data were downloaded from a server onto a student's device in a "pre-fetch" operation. This pre-fetch operation draws data from servers onto a student device in advance of the student seeing an item. This is done to avoid delays in presenting items to students. The pre-fetch operation may occur considerably in advance of when the student actually first sees an item. In the extreme case, all items associated with a performance task are pre-fetched at the same time, but a performance task may take a couple of hours for a student to complete.

Thus the pre-fetch timestamp does not help to determine the point in a test that a student experienced an interruption. This limitation makes several of the analyses originally planned impossible to conduct and generally weakens the usefulness of the results of this study. Grouping together all students that experienced any degree of interruption (regardless of how much of the test could have been affected by the interruption) likely dampens our ability to detect the effects of interruptions. This dampening is likely to occur because students who are interrupted near the end of their tests are likely to be affected negligibly, whereas those interrupted earlier are likely to be affected to a greater degree. The degree to which the grouping reduces the power of the analyses is unknown.

Apparent Inability to Identify Idiosyncratic Interruptions

In early discussions it appeared that students that experienced idiosyncratic interruptions could be identified by running one or more specific queries on the data tables. The data produced by those queries identified only students who had not yet completed logging in to the system. This conflicts with numerous reports from the field about idiosyncratic interruption events occurring during testing (not only during log in).

It is therefore possible that the number of students who experienced idiosyncratic interruptions is larger than the number of students who experienced system-wide interruption. Whether this is probable is also

[5] We use the word "apparently" and "appear" because documentation of the link between elements in data tables and specific events in the test administration system (as a whole) was insufficient to determine whether the necessary data exist. It is possible that such data do exist, but could not be identified.

unknown. If idiosyncratic interruptions do indeed have an effect on student test scores, being unable to identify who experienced them also likely dampens our ability to detect the effects of interruptions on test scores and on test completion/participation. Therefore, is less likely (to an unknown degree) that we can detect a true effect of interruptions because the following conditions *are reasonably likely* to exist in the available data:

- A large number of students who actually experienced an interruption are marked as not experiencing an interruption (i.e., those who experienced idiosyncratic interruptions).
- The interruptions did have an effect on those students' scores.
- The scores of the set of students that are marked as not interrupted will be contaminated to an unknown degree because the idiosyncratic interruptions could not be identified.

Table 3 presents the subsection of Table 2 the deals with server failure and server overloads, and highlights what data is available given the data limitations. In Table 2 the types of data that the support analyses of interruptions are given, with available data shown in black (A – interruption indicator, B – interruption start and stop timestamps) and unavailable data shown in red (C – interruption severity, D – item timestamps).

As can be seen in Table 3, available data are applicable to only one of the six types of interruptions intended to be investigated in this project. In addition, data supporting that investigation is limited, with only two of the three types of data available.

Table 3. *Data limitations of the project.*

	Type of Cause	
Type of Interruption	Server Failure	Server Overload
Involuntary log out	ABD	ABD
Inability to log in	ABD	ABD
Slow response & voluntary log out		ABCD
Slow response & continue working		ABCD
Scope of Interruption	**Systematic**	**Idiosyncratic**

ANALYSIS PLAN

The analysis plan was developed based on the framework displayed below in Figure 2. In the framework, after cleaning the data (step 1), three set of analyses were to be performed: on item-level data, on student-level data, and on aggregate-level data (i.e., school, district, and state).

For analyses of effects on item-level data (step 4A), it is necessary to identify whether each student responded to each item before, during or after a logout interruption. Without knowing when each item response was submitted, it is not useful to attempt to estimate effects on item scores. To be safe in such a situation, all item scores for all students that experienced a logout interruption must be treated as potentially affected (step 5A) and may need to be removed from later analyses, essentially excluding interrupted students from later analyses (step 6A).

For analyses of student-level data, as noted above it appears to be possible to construct a yes/no indicator of whether a student experienced an logout interruption from available data, but not at what point in a student's test the logout interruption occurred. This allows us to identify an overall effect of logout interruptions without regard to the severity of the logout interruptions (i.e., the number of items taken after the interruption). The apparent inability to identify idiosyncratic interruption further weakens our ability to quantify the effects of interruptions overall. Because it appears impossible to identify which students

experienced idiosyncratic interruptions, states might reasonably determine that they should treat all student scores from the 2015 spring administration as if they had been interrupted (e.g., apply steps 5B, 6B and 5C and 6C). It is possible to carry out analyses of the effects of experiencing (or not) a systematic interruption, but with the understanding that any effects are likely to be biased to an unknown degree (including the possibility of not detecting any effects when in fact there are effects).

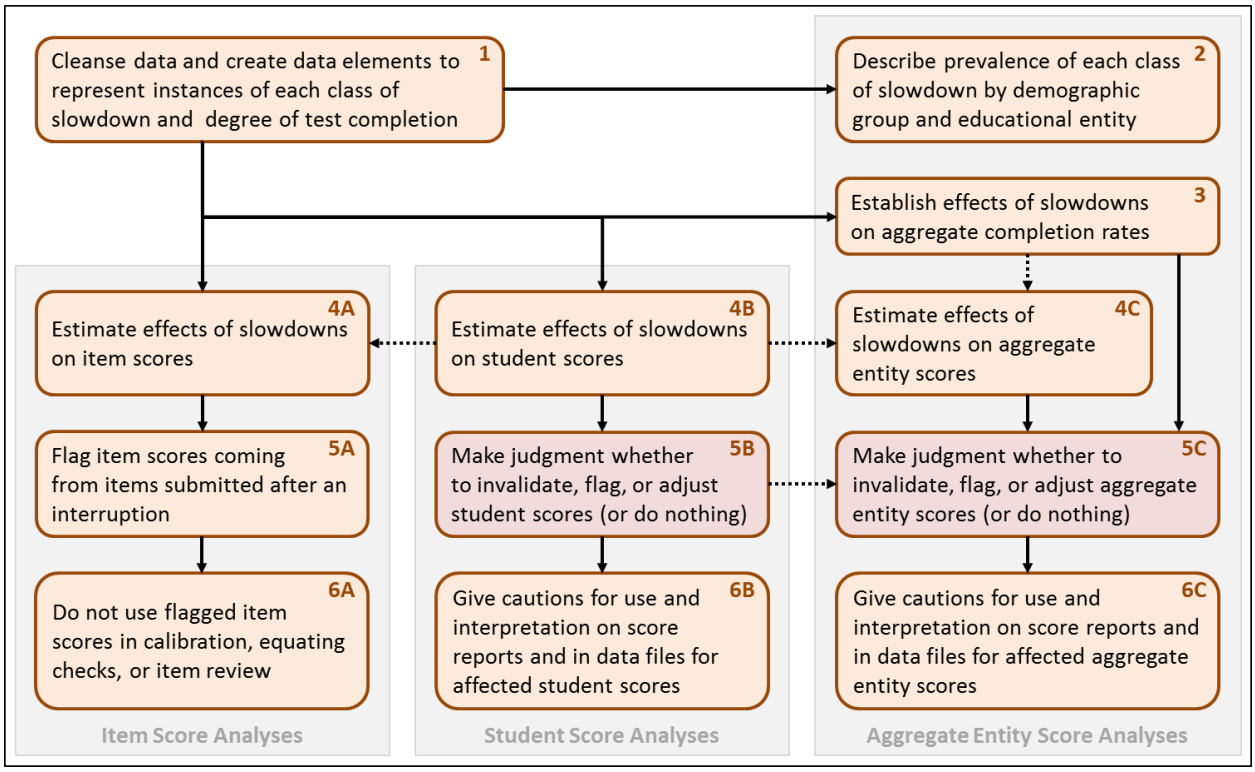

Figure 2. Framework for intended analyses.

For analyses of aggregate-level data, it is also important to have a clean delineation of which students experienced logout interruptions and to what degree those logout interruptions affected student scores. Neither of those criteria are met with currently available data. It is possible to carry out the analyses of the effects of experiencing (or not) a systematic logout interruptions on aggregate scores and on participation rates, but again with the understanding that those effects are likely to be understated to an unknown degree (including the possibility of not detecting any effects when in fact there are effects).

Given these limitations on the insights our studies can provide regarding the effects of interruptions, we now summarize two studies that are possible to conduct with currently available data. We then conclude with recommendations for addressing the data limitation for future administrations.

DESCRIPTIVE STATISTICS ON PREVALENCE OF LOGOUT INTERRUPTIONS

We first examine the prevalence of systematic logout interruptions resulting in involuntary logouts before quantifying the effects of those systematic logout interruptions. Recall that while there are numerous types of interruptions students may have experienced, the available data only captures students experiencing a systematic involuntary log out due to server failure. For simplicity we refer to these as "logout interruptions" throughout the description of the analyses and results of the study to avoid the inference that this analysis applies to anything other than the one type of interruption that was captured by the data.

The system failures that cased the logout interruptions occurred at different times in each state, as shown in Table 4. The events listed in Table 4 are derived from timestamped logs of students involuntarily logged out *en masse* when the server failed to respond in a timely manner, and corroborated by outages reported by Measured Progress.

Table 4. *Systematic involuntary logout events by state aggregated across grades (all times Eastern).*

Date	Montana			North Dakota			Nevada		
		# of Students			# of Students			# of Students	
	Timestamp	ELA	Math	Timestamp	ELA	Math	Timestamp	ELA	Math
April 14, 2015	10:45 AM	964	664	10:03 AM	952	379	11:45 AM	807	165
April 27, 2015	11:35 AM	1,108	747	11:35 AM	960	593	12:15 PM	3,946	1,923
	Total	2,072	1,411	**Total**	1,912	972	**Total**	4,753	2,088

Both ELA and mathematics were administered in two segments: a computer adaptive segment and a fixed-form performance task segment. In each state, some students experienced more than one logout interruption while testing. For example, one student was logged out twice while taking an ELA performance section in a single day, while another student was logged out while taking the Math multiple-choice section, then again on the following day while taking the ELA performance section. In Montana, 4 students were logged out more than once during their testing experience. In North Dakota, 51 students were logged out more than once and in Nevada, 73 students.

Table 5. *Total number of logout interruptions by state, subject, grade and segment.*

State	Grade	ELA		Math	
		Computer Adaptive	Performance Task	Computer Adaptive	Performance Task
Montana	3	203	74	297	29
	4	271	71	68	55
	5	276	92	71	35
	6	88	28	52	91
	7	286	61	315	103
	8	219	148	127	61
	11	90	165	47	60
North Dakota	3	184	67	113	7
	4	189	99	117	13
	5	179	50	41	103
	6	276	133	96	59
	7	272	23	102	18
	8	123	40	127	34
	11	165	112	44	98
Nevada	3	646	47	301	3
	4	639	19	357	43
	5	381	22	285	39
	6	540	151	129	54
	7	983	50	218	22
	8	902	373	413	224

Table 5 provides *counts of the number of logouts students experienced* while testing in each grade[6]. Students who experienced more than one logout are counted more than once in Table 5 below, e.g., the student who was logged out twice while taking the ELA performance section is counted twice in the cell for that section. The counts in Table 5, however, only show part of the picture. Table 6 and Figure 3 show the *number and percent of students who received scale scores* in each grade and subject combination who were interrupted while testing.

[6] Students who were logged out more than once are represented more than once in table 5.

Based on policy decisions by the states, students received a scale score only if they completed all sections of the assessment. Students who did not receive scale scores are excluded for our analyses. In Table 6 and the rest of the following analyses, we have flagged a student as experiencing a logout interruption if they were logged out during any section of the subject specific assessment. Across all three states, Figure 3 shows that a higher percentage of students experienced logout interruptions in ELA than in mathematics, with a few exceptions (grade in Montana, grade 8 in North Dakota, and grades 4-5 in Nevada). The percent of students with scores interrupted ranged from 0.6% in Nevada for grade-6 ELA to 5.8% in Nevada for grade-7 ELA.

Table 6. *Number and percent of students who received valid test scores and experienced any logout interruption by state, subject, and grade.*

Statistic	Logout Interrupted	State	Subject	Grade						
				3	4	5	6	7	8	11
Number of Students with Scores	No	MT	ELA	7,603	7,598	8,010	7,801	7,794	7,532	6,130
			Math	8,059	7,799	8,191	7,553	7,705	7,784	6,521
		ND	ELA	6,482	6,243	6,989	5,870	5,726	5,867	5,356
			Math	7,197	6,429	7,033	5,469	6,019	5,944	5,561
		NV	ELA	12,406	11,286	11,663	12,118	9,974	11,313	-
			Math	12,300	10,071	10,924	9,557	11,110	9,881	-
	Yes	MT	ELA	211	296	326	103	320	285	173
			Math	282	113	101	127	273	178	63
		ND	ELA	173	240	215	305	289	67	146
			Math	52	105	129	121	103	153	57
		NV	ELA	235	87	167	368	615	540	-
			Math	98	270	180	56	156	134	-
Percent of Students with Scores	Yes	MT	ELA	2.8	3.9	4.1	1.3	4.1	3.8	2.8
			Math	3.5	1.4	1.2	1.7	3.5	2.3	1.0
		ND	ELA	2.7	3.8	3.1	5.2	5.0	1.1	2.7
			Math	0.7	1.6	1.8	2.2	1.7	2.6	1.0
		NV	ELA	1.9	0.8	1.4	3.0	6.2	4.8	-
			Math	0.8	2.7	1.6	0.6	1.4	1.4	-

Finally, we computed the percent of students experiencing a logout interruption in each school and district in the dataset. Figure 4 show the distributions for districts and Figure 7 shows the distribution for schools. In both Figures 4 and 5, ELA is displayed in the top row with math in the bottom row; and the three columns represent Montana, North Dakota, and Nevada, respectively. The horizontal axes of the plots represent the percent of students interrupted in a district (figure 4) or school (figure 5). The vertical axis represents the percent of districts (figure 4) or schools (figure 5) that had the given percent of their students interrupted, or more. Thus, for the top left panel of Figure 4, it can be seen that at least 85% of schools in Montana had zero students who experienced a logout interruption in ELA (i.e., the lines in the graph start at 85% or above). Also, for example, in no Nevada district did more than approximately 80 percent of students experience a logout interruption in ELA (the lines reach only to approximately 80% in the top right panel) as compared to Montana and North Dakota in which at least one district had 100% of its students experience logout interruptions (the lines reach all the way to 100% in the top left and top middle panels).

As can be seen from both Figures 4 and 5, logout interruptions were more prevalent in ELA than in mathematics across schools, districts, and states. This is consistent with anecdotal reports of students taking the ELA assessment first, and thus experiencing the system-wide logout interruptions at occurred early in the assessment window.

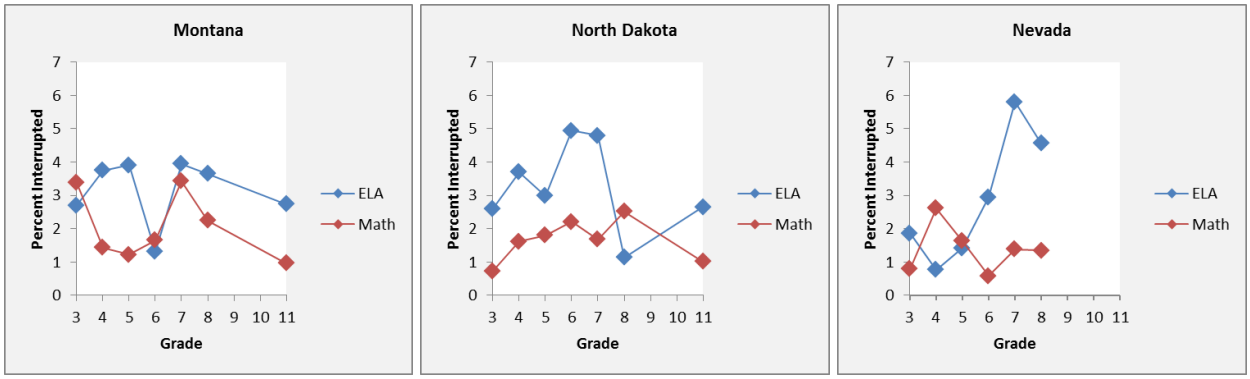

Figure 3. Percent of students with scores interrupted by state, grade, and subject area.

Figure 4 shows that about 15% or less of Montana or North Dakota districts had one or more students who experienced a logout interruption (in any grade in either subject- though other types of interruptions not captured by the data may have occurred). A larger percent of Nevada districts, 40%, had one or more students experiencing a logout interruption. This is in part due to the smaller number of districts in Nevada.

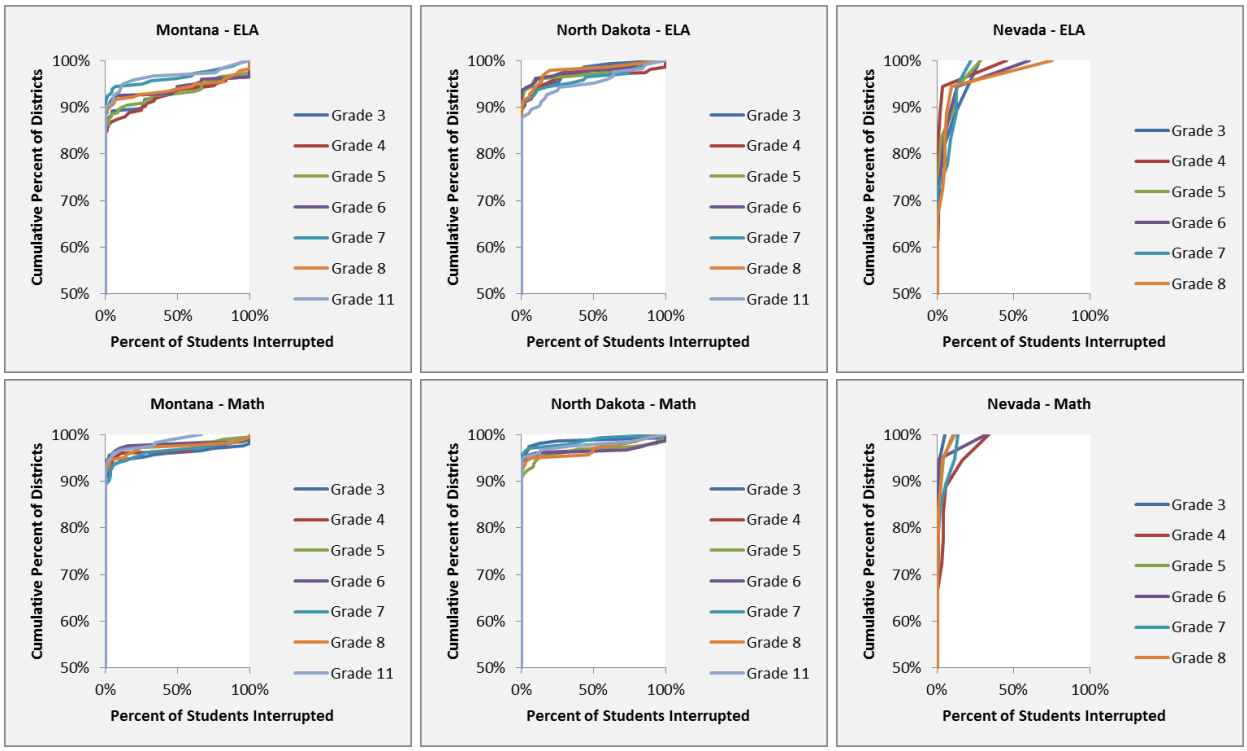

Figure 4. Distributions of logout interruption rates in *districts* across the three states.

Although the vast majority of Montana and North Dakota districts had zero students experience a logout interruption, the percent of students experiencing a logout interruption was 100% for some districts. By contrast the rate of logout interruptions reached approximately 80 percent in ELA and 40 percent for math in Nevada districts.

Figure 5 shows that about 12 percent or less of Montana or North Dakota schools had one or more students who experienced a logout interruption (in any grade in either subject- though other types of interruptions not captured by the data may have occurred). A larger percent of Nevada schools (about 20%) had one or more students experiencing a logout interruption in mathematics compared to approximately 10

percent in math. Although the vast majority of schools did not have any students experience a logout interruption, the logout interruption rates reached 100% for some schools in each of the three states.

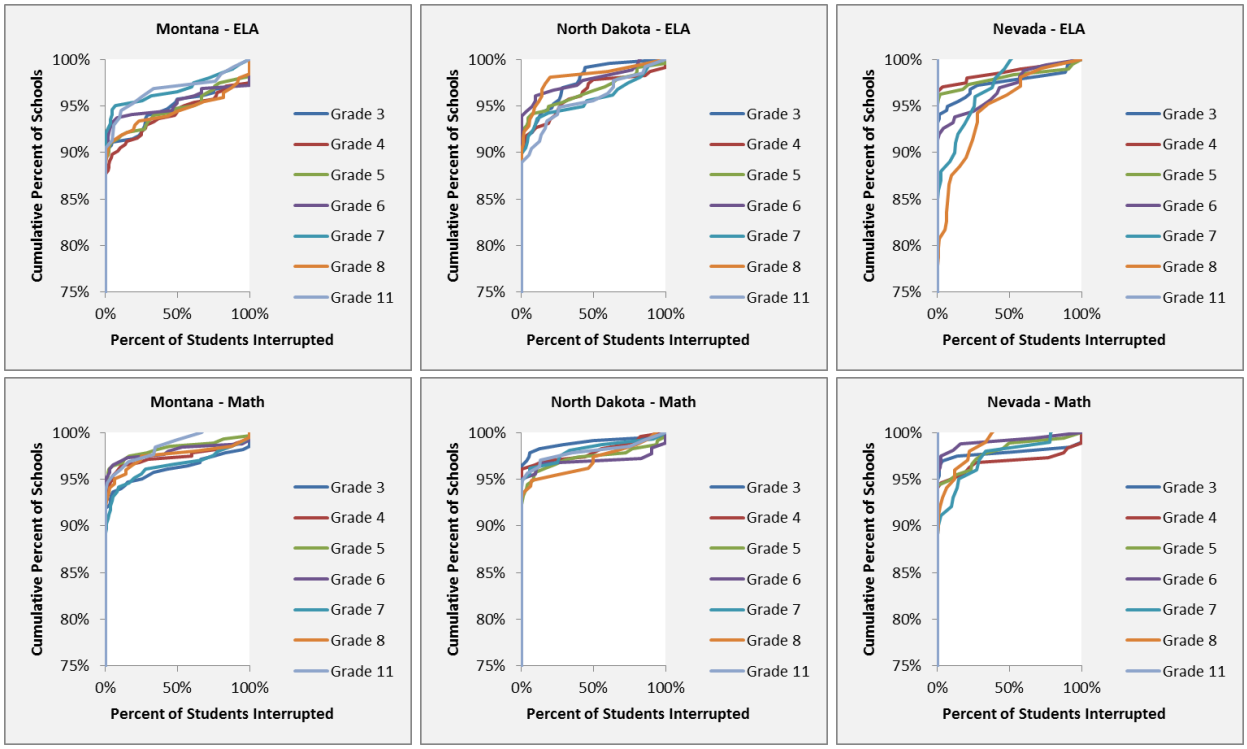

Figure 5. Distributions of logout interruption rates in *schools* across the three states.

Finally, we investigated the prevalence of logout interruptions by demographic group, as shown in Table 7. Table 7 shows the percent of students in each state and grade that experienced a logout interruption for various demographic groups considered protected, calculated using the entire group of students intended to take the test. The rows in Table 7 labeled "Rel. Prevalence" indicate how prevalent experiencing a logout interruption was for students in the protected group, relative to the prevalence for all students not in that group. When students in the protected group were more likely to experience a logout interruption, the cell is highlighted in red. When they were less likely to experience a logout interruption, the cell is highlighted in green. As can be seen from Table 7, there were considerable differences in the likelihood that students from the various demographic groups experienced a logout interruption:

- Relative prevalence of interruptions for minority students compared to their non-minority peers ranged from 0.56 times as prevalent (for North Dakota eleventh grade) to 1.34 times as prevalent (for North Dakota third grade).
- Relative prevalence for economically disadvantaged students compared to their non-disadvantaged peers ranged from 0.30 times as prevalent (for Nevada fifth grade) to 1.49 times as prevalent (for North Dakota third grade).
- Relative prevalence for students with disabilities compared to their non-disabled peers ranged from 0.32 times as prevalent (for Montana eleventh grade) to 1.26 times as prevalent (for North Dakota third grade).
- Relative prevalence for limited English proficient students compared to their English proficient peers ranged from 0.22 times as prevalent (for Montana fourth and eighth grades) to 2.36 times as prevalent (for North Dakota third grade).

Table 7. *Percent of students experiencing logout interruptions by demographic group, state, and grade.*

Demographic Variable	Group	Montana, Grade							North Dakota, Grade							Nevada, Grade					
		3	4	5	6	7	8	11	3	4	5	6	7	8	11	3	4	5	6	7	8
Minority	No	3.0	2.8	2.5	1.5	4.0	3.0	1.9	1.5	2.6	2.4	3.7	3.5	2.0	1.9	1.2	1.5	1.7	2.3	3.8	2.8
	Yes	3.2	1.6	2.8	1.2	2.1	2.6	1.5	2.0	3.1	2.5	3.2	2.1	1.1	1.0	1.5	1.9	1.3	1.3	3.2	3.4
	Rel. Prevalence	1.06	0.58	1.12	0.76	0.52	0.86	0.83	1.34	1.20	1.08	0.87	0.60	0.58	0.51	1.28	1.28	0.77	0.58	0.83	1.24
Economically Disadvantaged	No	3.5	2.6	2.8	1.5	3.8	2.9	2.0	1.4	2.4	2.4	3.9	3.5	1.9	1.8	1.6	1.5	2.2	2.3	4.3	2.9
	Yes	2.3	2.6	2.1	1.4	3.4	3.0	1.3	2.1	3.1	2.5	3.0	2.6	1.6	1.8	1.1	1.9	0.7	1.3	2.4	3.3
	Rel. Prevalence	0.66	0.98	0.74	0.91	0.91	1.02	0.66	1.49	1.26	1.04	0.77	0.73	0.85	0.97	0.67	1.29	0.30	0.57	0.56	1.15
Student with Disability	No	3.1	2.6	2.6	1.5	3.9	3.0	2.0	1.6	2.7	2.4	3.7	3.4	1.8	1.8	1.4	1.7	1.6	1.9	3.5	3.1
	Yes	2.7	2.3	2.4	1.4	2.3	2.3	0.6	2.0	2.5	2.6	3.3	2.4	2.0	2.1	0.9	1.0	1.1	1.7	3.4	2.8
	Rel. Prevalence	0.88	0.89	0.92	0.95	0.60	0.76	0.32	1.26	0.93	1.08	0.90	0.73	1.12	1.20	0.63	0.55	0.66	0.88	0.96	0.91
Limited English Proficient	No	3.1	2.7	2.5	1.5	3.7	3.0	1.8	1.5	2.6	2.4	3.6	3.2	1.8	1.9	1.3	1.5	1.6	2.0	3.6	3.1
	Yes	2.2	0.6	3.5	1.2	2.4	0.7	2.8	3.6	5.5	2.6	4.5	4.0	1.5	0.5	1.3	2.2	1.0	1.1	2.7	3.2
	Rel. Prevalence	0.71	0.22	1.38	0.80	0.65	0.22	1.54	2.36	2.15	1.10	1.23	1.25	0.84	0.26	1.02	1.45	0.64	0.53	0.74	1.04

STUDY 1: ESTIMATED EFFECTS OF LOGOUT INTERRUPTIONS ON STUDENT SCORES

Propensity Scores

The goal of the analyses in this section is to address the question "what score would a student who experienced a logout interruption have gotten, had he or she not experienced a logout interruption?" Like most investigations of this kind[7], we cannot answer this question directly, but instead investigate the *average* effect of logout interruptions on student scores. In this section we detail the approach we use to control for systematic differences between students who experienced logout interruptions and those who did not. In the next section we detail the model we use to estimate the average effect of logout interruptions. We investigate the effect of logout interruptions within each state for each grade and subject, meaning we repeat the analyses for each state, grade and subject combination.

In our estimation of the average logout effect, we use a propensity score approach to adjust for the non-random way in which students were exposed to logout interruptions. Specifically, we use logistic regression to estimate the propensity scores and include all of the student-level covariates available. However, differing circumstances between some schools and districts (e.g., staff reactions to logout interruptions) may have resulting in different chances of exposure to logout interruptions in those schools and districts relative to others. To investigate this possibility, we introduced random effects (i.e., a random intercept) for schools and districts into the propensity score model (c.f., Hong & Raudenbush, 2006; Kim & Seltzer, 2007). The variability of the random district intercepts was both statistically and practically insignificant, so we did not include random district intercepts in the final propensity scoring model.

Model. We use a multilevel (random effects) logistic regression to estimate the propensity scores. We modeled the probability that a student experienced a logout interruption as a function of a set of background covariates (*B*) and a random intercept for schools (τ_{0j}) with the following functional form:

$$\mathrm{P}(s_{ij}=1) = \exp\left(f(s_{ij})\right)/\left(1+\exp\left(f(s_{ij})\right)\right) \quad \text{Probability of experiencing a logout interruption}$$

$$f(s_{ij}=1) = \theta_{0j} + \sum_{b=1}^{B}\theta_b x_{bi} \quad \text{Student-level model (on logit scale)}$$

$$\theta_{0j} = \rho_{00} + \tau_{0j} \quad \text{School-level intercept model}$$

[7] A notable exception is the work of Sinharay et al. (2015), who were able to use the timing of an interruption to define "uninterrupted" and interrupted scores for each student in their study by matching uninterrupted students and interrupted students with similar characteristics and who were presented items in the same sequence.

where

$P(s_{ij} = 1)$ is the probability of student *i* in school *j* experiencing a logout slowdown,
$f(s_{ij} = 1)$ is the logit of *s,*
θ_b is the change in the log-odds of experiencing a logout interruption associated with a one standard deviation increase in one of six background variables,
ρ_{00} is the model intercept (i.e., grand mean), and
τ_{0j} is the random school effect for school *j*, capturing *j*'s deviation from the grand mean.

In combined form the model is:

$$f(s_{ij} = 1) = \rho_{00} + \tau_{0j} + \sum_{b=1}^{B} \theta_b x_{bi}$$

The background variables included in the model are (1) minority status (i.e., reported ethnicity is nonwhite), (2) educational disadvantage, (3) student disability status, (4) limited English proficiency, (5) standard normalized[8] prior[9] ELA test score and (6) standard normalized prior math test score. For third and/or fourth grade models[10], we did not have test scores from 2013-2014, as students are not tested in second grade.

These covariates are similar to, and in most cases the same as, those used in prior work examining interruptions (Bynum, Hoffman & Swain, 2013; Sinharay et al., 2104). These covariates are appropriate because they are known to associate strongly with the dependent variable, and they may associate with the "treatment" of experiencing a logout interruption. To ensure that these background variables do not bias the estimates of the effects of logout interruptions, they are included in the propensity score model.

Propensity scores were estimated using the R (R Core Team, 2015) package *lme4* (Bates et al, 2015). To validate the R code used, a subset of propensity scores was also estimated independently using HLM 7 (Raudenbush et al, 2011). Though the magnitude of the propensity scores differed between packages, the correlations were near perfect. The propensity scores produced using R were used for the remainder of the analyses.

Model of Effects of Logout Interruptions on Student Scores

The statistical model used to estimate average effects of experiencing a logout interruption is different than was originally proposed (see Appendix A for the originally proposed model). There were several reasons for this. First, the propensity scoring approach described above appears to only correct for bias in the main effect for logout interruptions. However, we proposed to model not just the main effect, but also differences in logout effects for various demographic groups (i.e., include interactions of demographics and logout interruptions). With only a single propensity score the interaction effects of logouts by demographic group may remain biased[11]. We used two approaches to address this issue.

[8] Normalized and standardized to mean = 0, variance = 1.
[9] The test score from the most recent state assessment the student took prior to 2015.
[10] Third grade models only for North Dakota and Nevada (as these states conducted state assessments in 2013-14). Third and fourth grade models for Montana (as state assessments were not conducted in 2013-14).
[11] For example, if there is no effect of a logout, but scores are generally lower for a specific demographic group, putting in an interaction term between logout interruptions and the demographic group would result in a negative effect for that demographic group by partially accounting for the generally lower scores of the group. Putting a main effect for the demographic group into the model ensures that the interaction term does not capture any part of a main effect of a demographic group.

First, we added the main effects of demographics into the model, in addition to the already present interaction terms. Propensity score matching to address selection bias is typically only performed for a single treatment effect, however the inclusion of the interaction terms means that, in essence, we have six treatment effects[12]. We are unaware of any application of propensity score matching to modeling both the main effect and interaction effects of a treatment. Adding the main effects can help to remove bias in the interaction effects, but this approach is not guaranteed to produce an unbiased effect for the interactions. Because of the potential weaknesses of the first approach, we also implemented an approach in which we created match groups of students and re-ran the model below using just the subset of students who experienced logout interruptions and their matched counterparts. To match students we used a Mahalanobis distance measure (Dodge, 2010) on the following demographics:

- Historically low-achieving minority status (non-white, non-Asian)
- Economically disadvantaged status
- Student with disability status
- Limited English proficiency
- Previous reading and math score[13]

Because there were relatively few students that experienced a logout interruption, finding exact matches was possible for the vast majority of interrupted students (i.e., identifying non-interrupted students who had exactly the same value of every matching variable). However, a few students that experienced logout interruptions had a unique set of background variables, and the nearest match for these students was selected. The multiple iterations were performed to assure that with many possible matches for the students with exact matches, sampling error was accounted for. This approach reduces the chance that the estimated effects of logout interruptions are unduly influenced by the sample of uninterrupted students used as matches. We compare the results of the two approaches. If the two models give considerably different results, that may signal potential issues with the models.

The model used for estimating the effect of logout interruptions using propensity score matching was the following linear regression model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * PS + \hat{\beta}_2 * min + \hat{\beta}_3 * ED + \hat{\beta}_4 * SWD + \hat{\beta}_5 * LEP + \hat{\beta}_6 * prior + \left(\hat{\beta}_7 + \hat{\beta}_8 * min + \hat{\beta}_9 * ED + \hat{\beta}_{10} * SWD + \hat{\beta}_{11} * LEP + \hat{\beta}_{12} * prior\right) * logout$$

Where

$\hat{y}$	is the predicted standard normalized 2015 Smarter Balanced ELA or math score,
PS	is the propensity score,
min	is a dummy variable for minority (non-White, non-Asian) status,
ED	is a dummy variable for economic disadvantage status,
SWD	is a dummy variable for student with disability status,
LEP	is a dummy variable for limited English proficient status,
$prior$	is the standard normalized score from the last state test given in the same subject as $\hat{y}$,

and the regression weights displayed in red are the effects of interest.

[12] Main effect of logout interruptions and interaction of logout interruptions with minority status, economically disadvantaged status, disability status, limited English proficient status, and previous score in the same subject.
[13] Students without prior scores were assumed to have a score at the population mean.

The model used for estimating the effect of logout interruptions based on the Mahalanobis distance matching with each iteration was the following (note that the propensity score is not used, as a reduced sample of students matched using the Mahalanobis distance was used in each iteration):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * min + \hat{\beta}_2 * ED + \hat{\beta}_3 * SWD + \hat{\beta}_4 * LEP + \hat{\beta}_5 * prior + \left(\hat{\beta}_6 + \hat{\beta}_7 * min + \hat{\beta}_8 * ED + \hat{\beta}_9 * SWD + \hat{\beta}_{10} * LEP + \hat{\beta}_{11} * prior\right) * logout$$

Again, the regression weights displayed in red are the effects of interest, and are directly analogous to those in the model using propensity scores. In both cases, because the standard normalized outcomes were used, and because dummy variables or standard normalized prior scores were used, the regression coefficients (for the propensity scored approach) and the mean regression coefficients (for the direct matching approach) can be interpreted as effect sizes.

No random district effects were included in the model because they did not account for significant variation. Random school effects were also not included because they also did not account for significant variation.

Results. Because both a main effect of logout interruptions and interactions with demographics (incremental effects of logout interruptions for students of different demographics) are included in the model, the results are presented as a baseline (main) effect and the main effect plus the incremental effect of logout interruptions for specific demographics. These results for the propensity scored model are shown in Table 8 and the average results for the 100 iterations of models using students matched on the Mahalanobis distance measure are presented in Table 9.

These tables bear some explanation. Both ELA and mathematics results are shown in these tables for each of the three states and each of the grades. The subject, state, and grade are shown down the left side of the tables. The regression coefficients from the model displayed above are provided in in the columns displayed in white. The baseline effect of logout interruptions is displayed in the first column displayed with colors. The overall (baseline plus incremental) effects of being interrupted for minority students (min), economically disadvantaged students (ED), students with disabilities (SWD), and limited English proficient students (LEP) are shown in the next four columns. The last four columns show the overall effects (baseline plus incremental) effects of being interrupted for students who scored very low, low, high, and very high on the most recent test in the same subject. Note that there are some cells with values of *NA*. These are for grades in which there was no prior score and for situations in which there were no LEP students that were interrupted. Positive effects of logout interruptions are displayed in green, and negative effects of logout interruptions are displayed in red, with larger effects in darker green or red. Finally, the partial row on the bottom right of each table shows the percentage of effects across subjects, states, and grades that were positive.

Table 8. *Results for model run using propensity score and all students.*

Subject	State	Grade	Regression coefficients													Logout effects								
			Main effects								Interactions of logout and						Base plus incremental effect of								
																						Prior z-score of…			
			Intercept	PS	Min	ED	SWD	LEP	Prior	Logout	Min	ED	SWD	LEP	Prior	Base	Min	ED	SWD	LEP	-2	-1	1	2
ELA	Montana	3	0.37	0.00	-0.59	-0.36	-0.72	-0.57	*NA*	0.05	-0.03	0.03	-0.17	0.92	*NA*	0.05	0.02	0.08	-0.12	0.97	*NA*	*NA*	*NA*	*NA*
		4	0.27	0.00	-0.51	-0.32	-0.86	-0.67	*NA*	0.08	0.11	-0.22	0.26	-0.25	*NA*	0.08	0.18	-0.15	0.33	-0.18	*NA*	*NA*	*NA*	*NA*
		5	0.03	-0.01	-0.21	-0.15	-0.28	-0.24	0.72	0.23	0.04	0.06	-0.09	-0.46	0.11	0.23	0.27	0.29	0.14	-0.23	0.01	0.12	0.34	0.45
		6	0.08	0.00	-0.18	-0.13	-0.30	-0.21	0.74	0.11	-0.32	0.00	-0.19	0.42	0.06	0.11	-0.21	0.11	-0.08	0.53	-0.02	0.05	0.17	0.23
		7	0.22	0.01	-0.15	-0.10	-0.31	-0.15	0.74	0.10	0.08	-0.15	-0.06	*NA*	0.08	0.10	0.17	-0.06	0.04	*NA*	-0.07	0.02	0.18	0.26
		8	0.23	0.02	-0.17	-0.13	-0.24	-0.30	0.73	0.02	0.21	0.04	-0.18	-0.28	-0.04	0.02	0.22	0.06	-0.17	-0.26	0.09	0.05	-0.02	-0.06
		11	0.01	0.00	-0.19	0.00	-0.27	-0.03	0.76	-0.14	0.37	0.00	0.30	0.41	0.06	-0.14	0.23	-0.14	0.16	0.27	-0.26	-0.20	-0.08	-0.02
	North Dakota	3	0.34	0.00	-0.27	-0.36	-0.53	-0.52	*NA*	-0.02	-0.08	0.44	-0.08	0.16	*NA*	-0.02	-0.11	0.41	-0.10	0.14	*NA*	*NA*	*NA*	*NA*
		4	0.09	-0.01	-0.08	-0.20	-0.08	-0.17	0.57	0.11	0.04	0.03	-0.01	-0.37	0.02	0.11	0.15	0.14	0.10	-0.26	0.07	0.09	0.13	0.15
		5	0.25	0.01	-0.08	-0.15	-0.19	-0.15	0.57	-0.10	0.19	-0.11	0.30	1.31	0.04	-0.10	0.09	-0.21	0.20	1.21	-0.17	-0.14	-0.06	-0.02
		6	0.17	0.01	-0.16	-0.16	-0.26	-0.29	0.56	0.12	-0.12	-0.09	-0.23	-0.11	-0.07	0.12	0.00	0.03	-0.11	0.01	0.26	0.19	0.05	-0.02
		7	0.04	-0.01	-0.07	-0.14	-0.32	-0.05	0.61	0.02	-0.20	0.16	0.04	0.08	0.01	0.02	-0.18	0.18	0.05	0.09	0.00	0.01	0.03	0.04
		8	0.18	0.00	-0.13	-0.17	-0.33	-0.16	0.59	0.06	0.04	-0.09	-0.24	0.95	0.07	0.06	0.10	-0.03	-0.18	1.01	-0.08	-0.01	0.13	0.20
		11	0.08	0.00	-0.07	-0.10	-0.32	-0.26	0.58	0.20	0.17	0.06	0.02	*NA*	-0.11	0.20	0.37	0.26	0.23	*NA*	0.43	0.32	0.09	-0.02
	Nevada	3	0.60	0.01	-0.27	-0.45	-0.83	-0.27	*NA*	-0.08	0.15	-0.22	0.35	-0.42	*NA*	-0.08	0.06	-0.31	0.26	-0.50	*NA*	*NA*	*NA*	*NA*
		4	0.19	0.01	-0.03	-0.17	-0.22	-0.09	0.77	-0.03	-0.24	0.50	0.31	-0.51	-0.20	-0.03	-0.27	0.47	0.28	-0.54	0.36	0.17	-0.22	-0.42
		5	-0.04	-0.01	-0.01	-0.13	-0.21	-0.07	0.80	0.20	-0.09	0.03	-0.26	-0.17	-0.03	0.20	0.11	0.23	-0.06	0.03	0.26	0.23	0.17	0.13
		6	0.13	0.01	-0.01	-0.09	-0.22	-0.02	0.77	-0.03	0.05	0.04	-0.06	0.16	0.01	-0.03	0.02	0.01	-0.09	0.13	-0.04	-0.03	-0.02	-0.02
		7	-0.01	0.00	-0.03	-0.05	-0.20	-0.11	0.84	0.10	0.06	-0.02	-0.08	-0.19	-0.06	0.10	0.17	0.08	0.02	-0.08	0.23	0.17	0.04	-0.02
		8	0.15	0.00	-0.07	-0.08	-0.29	-0.05	0.82	-0.07	-0.03	0.16	0.23	0.09	0.06	-0.07	-0.10	0.08	0.16	0.02	-0.19	-0.13	-0.02	0.04
Math	Montana	3	0.34	0.00	-0.63	-0.31	-0.78	-0.61	*NA*	-0.11	0.11	0.04	0.63	-0.18	*NA*	-0.11	0.01	-0.07	0.52	-0.28	*NA*	*NA*	*NA*	*NA*
		4	0.36	0.00	-0.60	-0.30	-0.82	-0.76	*NA*	0.13	-0.34	-0.30	0.00	0.04	*NA*	0.13	-0.21	-0.17	0.13	0.17	*NA*	*NA*	*NA*	*NA*
		5	-0.21	-0.03	-0.16	-0.16	-0.26	-0.26	0.78	0.25	-0.01	0.38	0.66	-0.22	0.22	0.25	0.24	0.63	0.92	0.03	-0.18	0.04	0.47	0.69
		6	0.09	0.00	-0.15	-0.09	-0.30	-0.19	0.80	0.14	-0.14	0.10	-0.11	0.36	0.05	0.14	0.00	0.24	0.03	0.50	0.04	0.09	0.18	0.23
		7	0.07	-0.01	-0.20	-0.07	-0.24	-0.01	0.81	0.11	-0.11	-0.08	-0.09	-0.16	-0.03	0.11	0.00	0.04	0.03	-0.05	0.17	0.14	0.08	0.05
		8	0.12	-0.01	-0.23	-0.11	-0.28	-0.25	0.82	0.17	0.18	-0.03	0.05	*NA*	0.02	0.17	0.34	0.14	0.22	*NA*	0.12	0.15	0.19	0.21
		11	-0.08	0.00	-0.14	-0.06	-0.19	-0.21	0.82	-0.11	-0.32	0.15	-0.22	*NA*	0.27	-0.11	-0.43	0.04	-0.33	*NA*	-0.65	-0.38	0.16	0.43
	North Dakota	3	0.51	0.01	-0.35	-0.38	-0.56	-0.48	*NA*	-0.35	-0.09	0.15	-0.20	-0.42	*NA*	-0.35	-0.44	-0.20	-0.55	-0.78	*NA*	*NA*	*NA*	*NA*
		4	0.22	0.00	-0.13	-0.23	-0.06	-0.12	0.64	-0.04	0.06	-0.09	-0.19	-0.30	-0.07	-0.04	0.02	-0.13	-0.24	-0.35	0.10	0.03	-0.11	-0.18
		5	0.10	-0.01	-0.11	-0.17	-0.13	0.00	0.64	0.21	-0.44	-0.05	-0.03	-0.29	-0.32	0.21	-0.23	0.17	0.19	-0.07	0.86	0.54	-0.11	-0.43
		6	0.15	0.00	-0.18	-0.18	-0.22	-0.28	0.62	0.25	-0.10	-0.04	0.26	0.16	-0.06	0.25	0.16	0.22	0.52	0.41	0.37	0.31	0.20	0.14
		7	0.11	-0.01	-0.12	-0.12	-0.25	-0.08	0.67	0.00	0.21	-0.13	0.04	*NA*	-0.06	0.00	0.21	-0.13	0.04	*NA*	0.13	0.06	-0.06	-0.13
		8	0.37	0.01	-0.14	-0.19	-0.22	-0.03	0.69	-0.12	-0.11	0.00	-0.06	-0.27	-0.04	-0.12	-0.22	-0.12	-0.17	-0.38	-0.04	-0.08	-0.15	-0.19
		11	0.26	0.01	-0.09	-0.11	-0.31	-0.12	0.65	-0.01	0.29	-0.31	-0.14	*NA*	0.13	-0.01	0.28	-0.32	-0.15	*NA*	-0.28	-0.15	0.12	0.25
	Nevada	3	0.53	0.01	-0.30	-0.43	-0.76	-0.21	*NA*	-0.43	0.05	0.05	-0.38	0.11	*NA*	-0.43	-0.37	-0.37	-0.80	-0.32	*NA*	*NA*	*NA*	*NA*
		4	0.10	0.00	-0.11	-0.17	-0.27	-0.06	0.76	-0.12	0.19	0.09	0.14	0.15	0.08	-0.12	0.07	-0.03	0.02	0.03	-0.27	-0.20	-0.04	0.03
		5	-0.08	-0.01	-0.05	-0.09	-0.26	-0.09	0.82	0.16	0.02	0.04	0.06	-0.06	-0.01	0.16	0.19	0.20	0.22	0.10	0.19	0.18	0.15	0.14
		6	0.22	0.01	-0.08	-0.10	-0.30	-0.03	0.83	0.25	-0.41	-0.24	-0.53	0.33	-0.22	0.25	-0.16	0.01	-0.28	0.57	0.68	0.46	0.03	-0.18
		7	-0.02	0.01	-0.06	-0.07	-0.12	-0.07	0.87	0.09	-0.08	0.04	0.06	-0.25	-0.07	0.09	0.01	0.13	0.15	-0.16	0.23	0.16	0.01	-0.06
		8	0.07	0.00	-0.07	-0.07	-0.17	-0.05	0.85	0.31	-0.01	-0.08	-0.66	-0.10	-0.03	0.31	0.30	0.22	-0.36	0.21	0.37	0.34	0.27	0.24
											Percent positive effects					60.0	67.5	62.5	60.0	55.9	59.4	71.9	65.6	56.3

As can be seen from these tables, the results are similar (though not identical) when using either model (propensity scoring for Table 8, Mahalanobis matching for Table 9). In addition, positive effects of logout interruptions outnumber negative effects of logout interruptions regardless of the model used or the effect in question (baseline or baseline plus increment). However, while the pattern of more positive effects than negative effects is clear, there are also a considerable number of negative effects of logout interruptions (ranging from 28 to 50 percent, depending on the model used, state, subject, grade, and the effect in question). The effects tended to be most extreme with LEP students. However the number of LEP students who experienced logout interruptions tended to have the smallest sample size of any demographic group studied. The extreme effects may be real effects of logout interruptions for these students, or they may be more variable because of the relatively small sample sizes. The effects also tend to be more extreme for students with disabilities, but they also vary between negative and positive effects.

Table 9. *Mean results for models run using 100 iterations of Mahalanobis matched students.*

Subject	State	Grade	Regression coefficients												Logout effects								
			Main effect							Interactions of Logout and						Base plus incremental effect of							
																				Prior z-score of…			
			Intercept	Min	ED	SWD	LEP	Prior	Logout	Min	ED	SWD	LEP	Prior	Base	Min	ED	SWD	LEP	-2	-1	1	2
ELA	Montana	3	0.32	-0.70	-0.30	-0.69	-0.12	*NA*	0.11	0.07	-0.03	-0.20	0.48	*NA*	0.11	0.18	0.07	-0.09	0.58	*NA*	*NA*	*NA*	*NA*
		4	0.29	-0.48	-0.33	-0.89	-0.28	*NA*	0.05	0.08	-0.21	0.29	-0.64	*NA*	0.05	0.14	-0.16	0.35	-0.58	*NA*	*NA*	*NA*	*NA*
		5	0.14	-0.21	-0.24	-0.43	-0.23	0.77	0.13	-0.04	0.17	0.04	-0.44	0.05	0.13	0.09	0.31	0.17	-0.30	0.02	0.08	0.19	0.24
		6	0.08	-0.32	-0.16	-0.58	-0.18	0.68	0.07	-0.13	-0.01	0.09	0.47	0.12	0.07	-0.07	0.06	0.15	0.54	-0.17	-0.05	0.18	0.30
		7	0.14	-0.19	-0.19	-0.46	-0.41	0.80	0.25	0.08	-0.15	-0.08	-1.25	-0.02	0.25	0.33	0.10	0.17	-0.99	0.29	0.27	0.24	0.22
		8	0.03	-0.15	-0.13	-0.40	-1.11	0.75	0.22	0.16	0.02	-0.09	0.47	-0.07	0.22	0.37	0.23	0.12	0.68	0.36	0.29	0.14	0.07
		11	0.13	-0.27	-0.10	-0.41	0.16	0.69	-0.27	0.43	0.07	0.47	0.58	0.13	-0.27	0.16	-0.20	0.20	0.31	-0.52	-0.40	-0.15	-0.02
	North Dakota	3	0.35	-0.27	-0.33	-0.58	-0.62	*NA*	-0.04	-0.08	0.41	-0.02	0.26	*NA*	-0.04	-0.12	0.38	-0.06	0.22	*NA*	*NA*	*NA*	*NA*
		4	0.23	-0.09	-0.32	-0.43	-0.51	0.58	-0.04	-0.02	0.24	0.29	0.10	0.00	-0.04	-0.05	0.20	0.25	0.07	-0.04	-0.04	-0.04	-0.04
		5	0.17	-0.38	-0.06	-0.60	0.00	0.54	0.01	0.38	-0.19	0.57	1.18	0.04	0.01	0.38	-0.19	0.57	1.18	-0.08	-0.04	0.05	0.09
		6	0.17	-0.23	-0.22	-0.72	-0.68	0.54	0.12	-0.04	-0.07	0.15	0.21	-0.07	0.12	0.08	0.06	0.28	0.33	0.25	0.19	0.06	-0.01
		7	0.17	-0.04	-0.20	-0.76	-0.94	0.58	-0.05	-0.11	0.09	0.34	0.46	-0.02	-0.05	-0.16	0.04	0.29	0.41	0.00	-0.02	-0.07	-0.10
		8	0.16	-0.07	-0.36	-0.65	-0.63	0.59	0.16	0.08	-0.04	-0.11	0.01	-0.12	0.16	0.23	0.11	0.05	0.17	0.39	0.27	0.04	-0.07
		11	0.10	-0.27	-0.32	-0.89	-0.72	0.57	0.23	0.52	0.22	0.24	*NA*	-0.12	0.23	0.75	0.46	0.47	*NA*	0.47	0.35	0.11	0.00
	Nevada	3	0.46	-0.26	-0.46	-0.84	-0.27	*NA*	0.05	0.15	-0.20	0.36	-0.41	*NA*	0.05	0.19	-0.15	0.41	-0.36	*NA*	*NA*	*NA*	*NA*
		4	0.25	-0.04	-0.28	-0.43	-0.15	0.72	-0.12	-0.22	0.51	0.56	-0.31	-0.13	-0.12	-0.33	0.39	0.44	-0.43	0.14	0.01	-0.25	-0.38
		5	0.17	-0.13	-0.07	-0.44	-0.32	0.75	0.06	-0.08	-0.01	-0.26	0.05	0.00	0.06	-0.02	0.04	-0.20	0.11	0.07	0.06	0.06	0.05
		6	0.14	-0.03	-0.16	-0.58	-0.41	0.71	0.03	0.05	-0.01	0.14	0.25	0.02	0.03	0.08	0.02	0.17	0.27	0.00	0.01	0.04	0.06
		7	0.09	-0.04	-0.14	-0.45	-0.31	0.78	0.07	0.05	0.04	-0.01	-0.05	-0.04	0.07	0.12	0.11	0.06	0.02	0.14	0.11	0.04	0.00
		8	0.28	-0.17	-0.18	-0.58	-0.47	0.74	-0.01	-0.03	0.13	0.22	0.16	0.01	-0.01	-0.04	0.12	0.21	0.16	-0.04	-0.02	0.01	0.02
Math	Montana	3	0.33	-0.56	-0.31	-0.77	-0.71	*NA*	-0.10	0.04	0.04	0.62	-0.08	*NA*	-0.10	-0.06	-0.05	0.52	-0.18	*NA*	*NA*	*NA*	*NA*
		4	0.31	-0.62	-0.35	-0.79	-0.68	*NA*	0.19	-0.32	-0.25	-0.04	-0.04	*NA*	0.19	-0.13	-0.07	0.15	0.15	*NA*	*NA*	*NA*	*NA*
		5	0.15	-0.27	-0.13	-0.42	-0.25	0.80	-0.08	0.14	0.32	0.78	-0.29	0.18	-0.08	0.06	0.24	0.70	-0.36	-0.44	-0.26	0.11	0.29
		6	0.12	-0.29	-0.11	-0.26	-0.08	0.82	0.12	0.16	0.16	-0.21	0.14	0.02	0.12	0.29	0.28	-0.08	0.26	0.09	0.10	0.14	0.15
		7	0.12	-0.19	-0.11	-0.28	-0.33	0.84	0.08	-0.14	-0.02	-0.05	-0.03	-0.07	0.08	-0.06	0.07	0.03	0.05	0.21	0.15	0.02	-0.05
		8	0.28	-0.38	-0.13	-0.32	-0.19	0.79	-0.01	0.14	0.08	-0.03	*NA*	0.04	-0.01	0.14	0.08	-0.03	*NA*	-0.08	-0.04	0.03	0.07
		11	0.01	-0.04	-0.06	-0.37	-0.09	0.85	-0.14	-0.46	0.15	-0.20	*NA*	0.20	-0.14	-0.60	0.01	-0.34	*NA*	-0.55	-0.34	0.06	0.26
	North Dakota	3	0.30	-0.25	-0.42	-0.44	-0.46	*NA*	-0.14	-0.19	0.21	-0.31	-0.45	*NA*	-0.14	-0.33	0.06	-0.45	-0.59	*NA*	*NA*	*NA*	*NA*
		4	0.25	-0.35	-0.24	-0.30	-0.22	0.57	-0.05	0.24	-0.07	-0.03	-0.45	-0.04	-0.05	0.19	-0.12	-0.08	-0.50	0.02	-0.02	-0.09	-0.12
		5	0.24	-0.07	-0.35	-0.38	-0.26	0.64	0.05	-0.51	0.15	0.20	-0.06	-0.33	0.05	-0.46	0.20	0.25	-0.01	0.70	0.37	-0.28	-0.60
		6	0.21	-0.41	-0.30	-0.56	-0.78	0.62	0.21	0.29	-0.06	0.14	0.48	-0.07	0.21	0.50	0.16	0.35	0.70	0.36	0.28	0.14	0.07
		7	0.17	-0.41	-0.03	-0.49	0.04	0.70	-0.01	0.33	-0.22	0.13	0.21	-0.11	-0.01	0.32	-0.24	0.11	0.20	0.21	0.10	-0.12	-0.24
		8	0.33	-0.33	-0.24	-0.56	-0.08	0.71	-0.10	0.01	0.05	0.24	-0.15	-0.05	-0.10	-0.08	-0.05	0.15	-0.24	0.01	-0.04	-0.15	-0.21
		11	0.11	-0.35	-0.09	-0.28	-1.05	0.75	0.21	0.61	-0.39	-0.40	0.78	-0.07	0.21	0.82	-0.18	-0.19	1.00	0.36	0.29	0.14	0.06
	Nevada	3	0.42	-0.28	-0.40	-0.71	-0.24	*NA*	-0.31	0.03	0.04	-0.43	0.15	*NA*	-0.31	-0.28	-0.28	-0.74	-0.17	*NA*	*NA*	*NA*	*NA*
		4	0.10	-0.16	-0.20	-0.49	0.03	0.76	-0.08	0.11	0.10	0.17	0.11	0.03	-0.08	0.03	0.03	0.09	0.04	-0.14	-0.11	-0.04	-0.01
		5	0.08	-0.13	-0.19	-0.40	-0.05	0.83	0.06	-0.01	0.08	0.12	-0.14	-0.05	0.06	0.06	0.14	0.19	-0.07	0.16	0.11	0.02	-0.03
		6	0.17	-0.01	-0.27	-0.61	-0.21	0.80	0.24	-0.52	-0.07	-0.33	-0.12	-0.20	0.24	-0.27	0.17	-0.08	0.12	0.65	0.45	0.04	-0.16
		7	-0.07	-0.12	-0.03	-0.43	-0.07	0.89	0.16	0.03	-0.08	0.31	-0.25	-0.12	0.16	0.19	0.08	0.47	-0.09	0.40	0.28	0.04	-0.08
		8	0.09	-0.13	-0.13	-0.10	-0.27	0.82	0.31	0.06	-0.05	-0.78	0.03	-0.02	0.31	0.36	0.26	-0.48	0.34	0.36	0.33	0.28	0.26
										Percent positive effects					60.0	60.0	72.5	70.0	62.2	68.8	62.5	71.9	50.0

To evaluate other potential patterns in the data, the marginal percentages of effects that are positive were calculated for each state, subject, and grade. These are shown in Table 10 where it can be seen that there is disagreement about effects in mathematics being generally less positive than in ELA (depending on method used to estimate effects). There was a slightly lesser preponderance of positive effects in North Dakota than in other states. Positive effects were in the majority in grades 5-8, but not in grades 3-4, with conflicting results for whether positive or negative effects were in the majority in grade 11 (again depending on method used to estimate effects).

The size of the effects is an important part of understanding how important the effects of logout interruptions are. Because the effects vary from negative to positive, we present a distribution of effects for each potential effect of logout interruptions. We classify effect sizes in the following manner:

- Large negative when $\beta \leq -0.5$ (denoted in the figures as – – –)
- Moderate negative when $-0.5 < \beta \leq -0.3$ (denoted in the figures as – –)
- Small negative when $-0.3 < \beta \leq -0.1$ (denoted in the figures as –)
- Negligible when $-0.1 < \beta < 0.1$ (denoted in the figures as ±)
- Small positive when $0.1 \leq \beta < 0.3$ (denoted in the figures as +)
- Moderate positive when $0.3 \leq \beta < 0.5$ (denoted in the figures as ++)
- Large positive when $0.5 \leq \beta$ (denoted in the figures as +++)

Table 10. *Marginal percentages of positive effects of logout interruptions by subject, state, and grade.*

Model	Subject		State			Grade						
	ELA	Math	MT	ND	NV	3	4	5	6	7	8	11
Propensity Scoring	63.0	61.3	69.4	54.3	62.0	18.5	46.3	75.9	72.2	72.2	55.6	44.4
Mahalanobis Matching	70.6	58.0	65.5	59.5	68.0	20.4	35.2	70.4	83.3	74.1	66.7	55.6

Our thresholds for small, moderate, and large effect sizes are conservative, given that thresholds tend to be rules of thumb and the stakes attached to student test scores tend to be high. These distributions (the percent of effect sizes in each category across states, grades, and subjects) shown in Figures 6-10. In each of these figures, distributions of effect sizes are displayed for both the propensity scoring model (light blue) and Mahalanobis matching model (dark blue). In all of these figures, panels for ELA are in the top row with panels for math in the bottom row.

As shown in Figures 6-10, the results of the model using propensity scores and the model using mean results of 100 iterations of Mahalanobis-matched samples were quite similar with a small tendency toward more positive effects with Mahalanobis matching, providing some assurance that the results are not an artifact of the matching method employed.

As seen in Figure 6, the base effect of experiencing a logout interruption (for non-SWD, non-LEP, non-minority students with average prior scores) ranged from small negative to small positive in ELA and from moderate negative to moderate positive in math, with a slight negative shift of the effects in math compared to the effects in ELA.

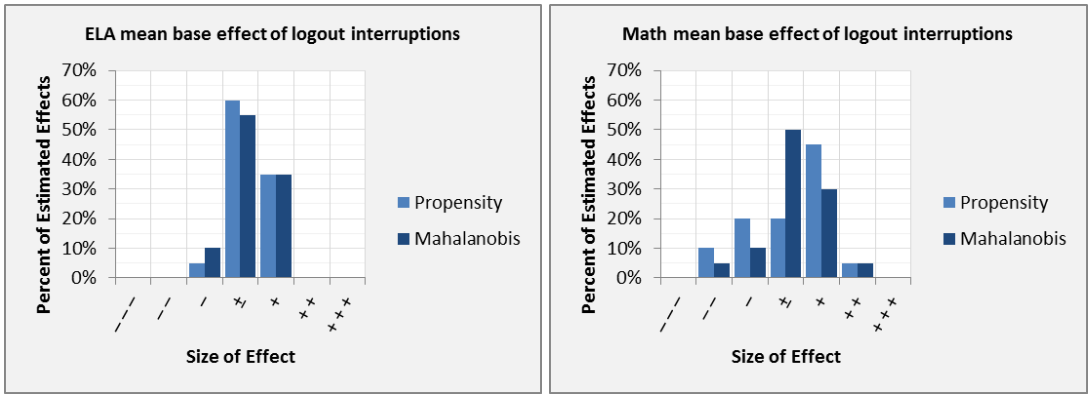

Figure 6. Distribution of mean base effect of logout interruptions across states and grades.

The effects tend to become more dispersed when evaluated on the basis of demographics. As shown in Figures 7-8, effect sizes range from large negative to large positive for all but economically disadvantaged students where the effects range from moderate negative to large positive. In addition, these figures show that there is a modest pattern of more disperse (and possibly more negative) effects on math scores compared to ELA.

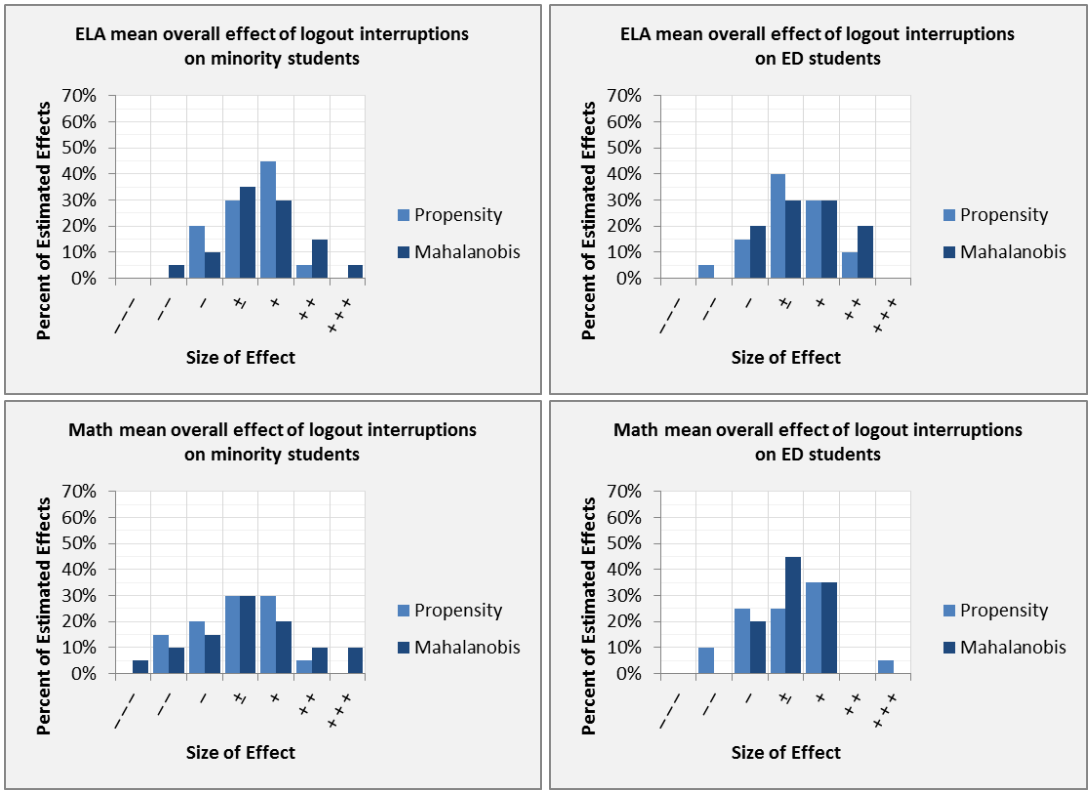

Figure 7. Distribution of effects of logout interruptions for minority and ED demographic groups.

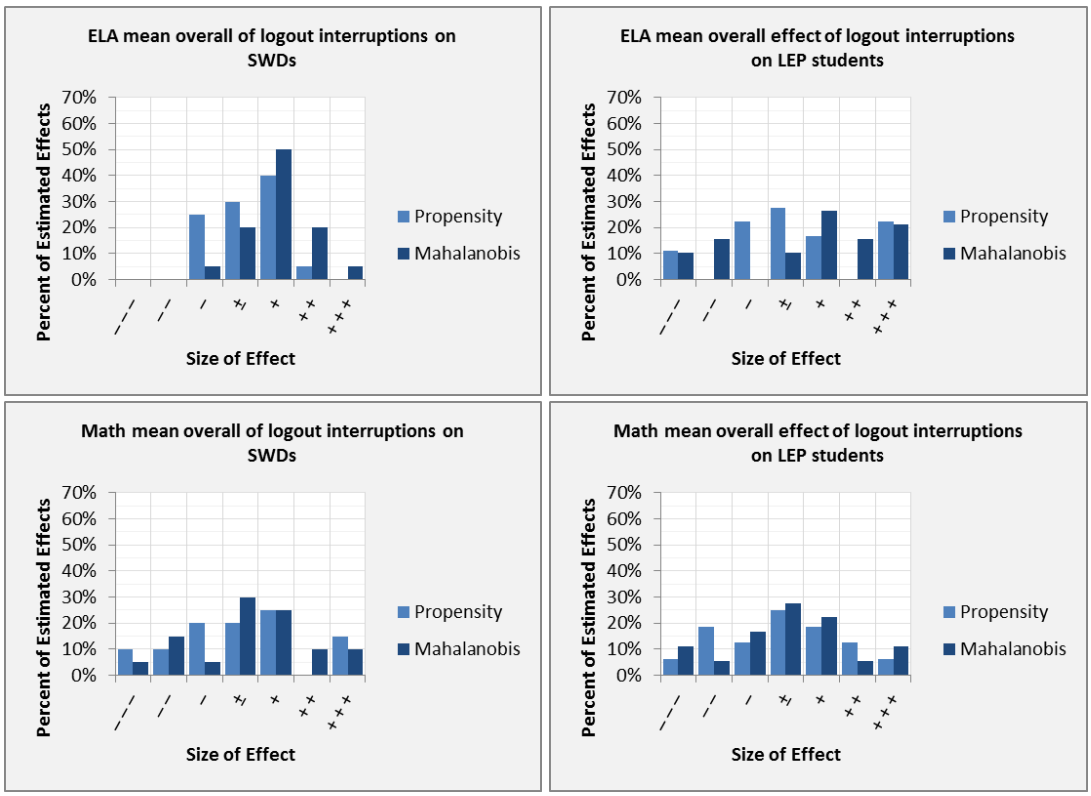

Figure 8. Distribution of effects of logout interruptions for SWD and LEP demographic groups.

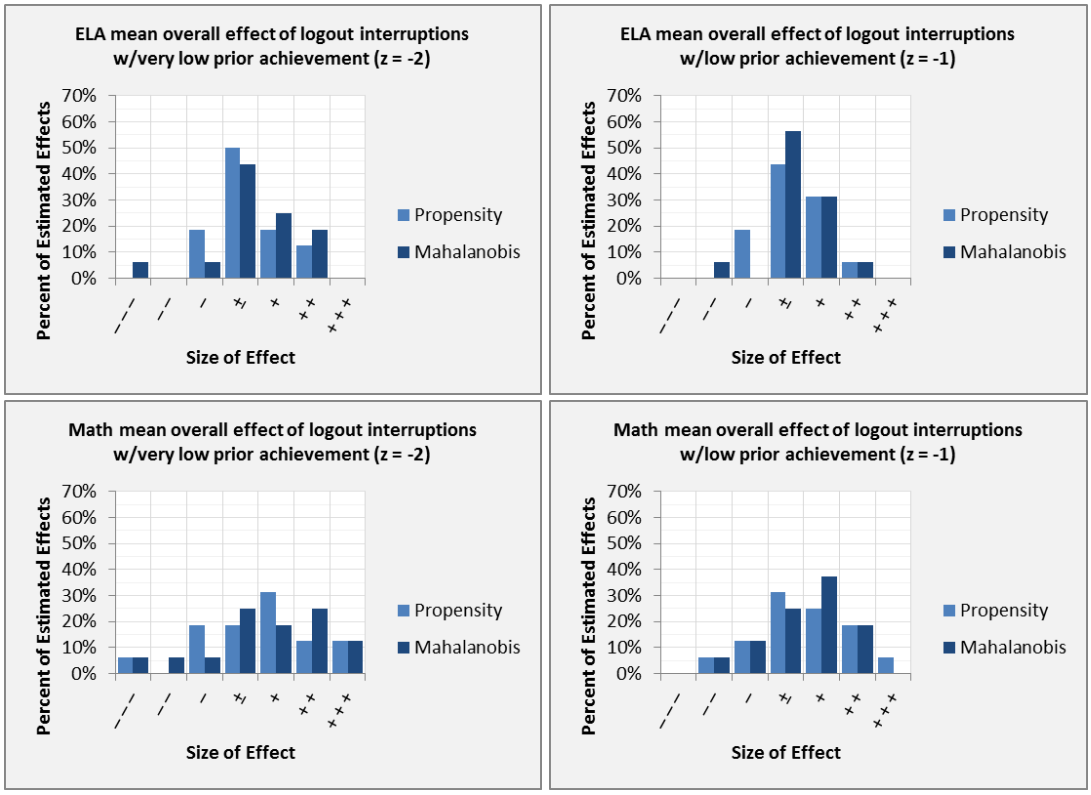

Figure 9. Distribution of effects of logout interruptions for of very low and low prior achievement levels.

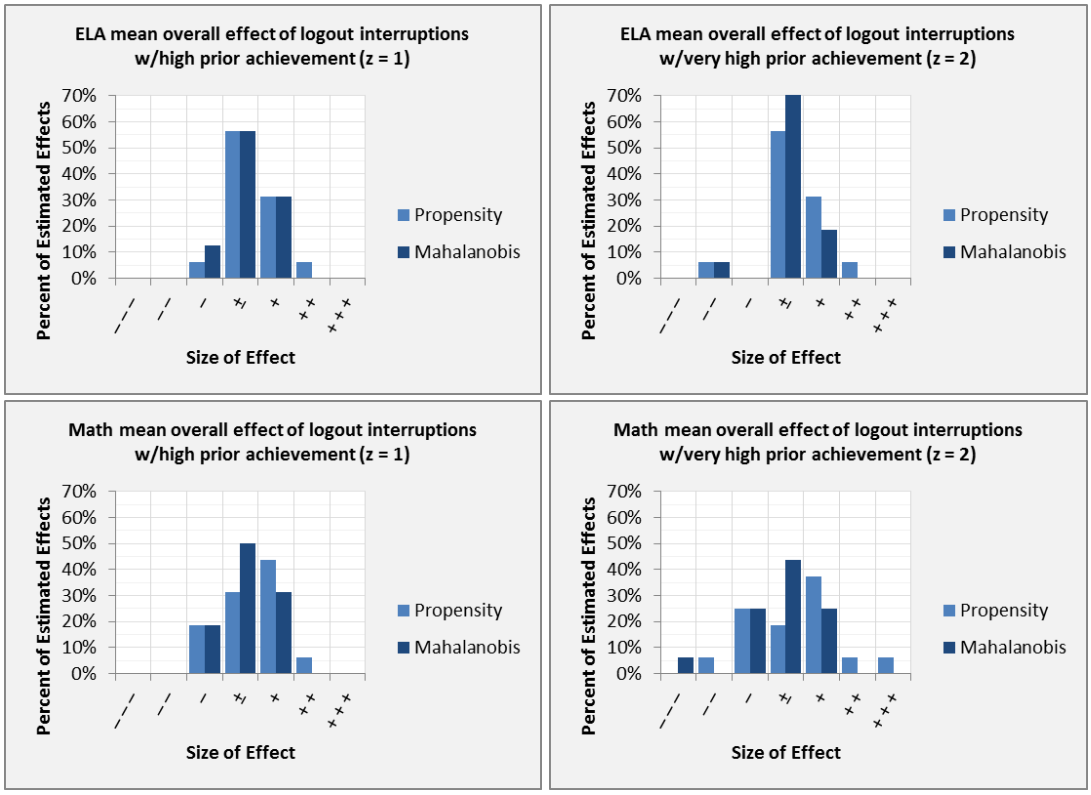

Figure 10. Distribution of effects of logout interruptions for high and very high prior achievement levels.

Figures 9-10 shows the distributions of overall effects of logout interruptions for students with very low, low, high, and very high prior test scores in the same subject (z-scores of -2, -1, 1, and 2, respectively).

These effects also range from large negative to large positive very low and very high previous scoring students and from moderate negative to moderate positive for low and high previous scoring students. Again, these figures show a modest pattern of more disperse (and for Mahalanobis matching, more negative effects) on math scores.

Discussion. In general, effects of logout interruptions were mostly small or negligible effects with some moderate to large negative and positive effects.

There tended to be more positive effects of logout interruptions than negative effects. Relative to ELA, the effects of logout interruptions in mathematics were more variable (i.e., more dispersed). In addition, there were more negative effects in mathematics than ELA, particularly when using one of the two methods for matching interrupted and non-interrupted students. There also were slightly fewer positive effects in North Dakota relative to the other states. Finally, positive effects were more numerous than negative effects in grades 5-8, whereas there were more negative effects in grades 3-4. In grade 11, the prevalence of negative over positive effects depended on the method used for matching interrupted and non-interrupted students.

While there were some patterns as described above, the patterns tended to be weak. That is, there were a substantial number of both negative and positive effects, even when there was a prevalence of negative or positive effects. This lack of strength in patterns could reflect the true state of affairs. However, it could also be the case that the results do not show clear patterns because:

- We had a blunt marker of experiencing a logout interruption (we were only able to identify whether a student experienced a logout interruption, not what proportion of her test the logout interruption could be expected to affect). Because of relatively small sample sizes, this may have resulted in chance identification of some effects, diluting the patterns in the data.
- We were only able to identify students who experienced one of the six types of interruptions that we anticipate being able to study. This may have resulted in noisy data in which students who experienced interruptions equally important to the type we could capture not being identified, again, diluting the patterns in the data.

STUDY 2: DIFFERENCES IN PERCENT OF STUDENTS RECEIVING TEST SCORES COMPARED TO PREVIOUS YEARS

In all three states, schools were allowed significant flexibility to determine whether to continue testing in light of the interruptions. To investigate *potential* effects of logout interruptions on the percent of students receiving test scores, baseline participation data from 2011-12, 2012-13, and 2013-14 were obtained from Montana and North Dakota[14]. Using these data, the weighted three-year average participation rate (or rate of students receiving valid test scores) for each district and school was calculated. The percent of students receiving valid test scores in 2015 was also calculated. The difference between the three-year weighted average proportion with valid test scores and the 2015 proportion of students with valid test scores was tested for statistical significance at the $\alpha = 0.05$, 0.10, and 0.20 levels. Changes in proportions receiving valid test scores were identified as showing a significant decrease, a significant increase, or a non-statistically significant change. The proportions of the schools and districts with a significant decrease and significant increase in rates of students with valid test scores are shown in Table 11. These differences suggest, but do not definitively establish, that logout interruptions resulted in a decrease in test completion and/or participation rates.

[14] Data from Nevada were unavailable.

Along the left-most column of Table 11 is the level of confidence used to identify statistically significant changes in rates of receiving valid scores. Three levels of confidence ($\alpha = 0.05$, 0.10, and 0.20) were used to provide states with identifications of schools and districts with changes in proportion of students receiving valid test scores that are very likely, likely, and somewhat likely related to the existence of logout interruptions in spring 2015. From the standpoint of identifying an overall relationship between of logout interruptions in 2015 and rates of students receiving valid scores in schools or districts, we use the $\alpha = 0.05$ confidence level. As can be seen from Table 11, using the confidence level of $\alpha = 0.05$, there was no state/grade/subject combination with greater than 5 percent of either schools or districts with a statistically significant increase in the proportion of students receiving valid scores; in fact, the vast majority of combinations had zero percent of schools and districts with statistically significant increases. On the other hand, there was no combination with less than 15 percent of either schools or districts experiencing a statistically significant decrease in the proportion of students receiving valid test scores; in fact, the vast majority of combinations were above 20 percent.

Table 11. *Districts/schools with a decrease/increase in proportion of students with valid test scores by state, grade, and subject.*

Confidence	Grade	Percent w/Significant Decrease in Participation								Percent w/Significant Increase in Participation							
		Montana				North Dakota				Montana				North Dakota			
		Districts		Schools		Districts		Schools		Districts		Schools		Districts		Schools	
		ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math
α = 0.05	3	36.2	21.7	38.4	17.6	26.8	25.0	27.0	27.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	4	34.1	29.6	35.2	29.4	22.5	25.4	22.8	26.8	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
	5	20.5	18.8	21.8	17.7	25.4	21.3	24.4	19.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	6	27.4	37.6	29.2	37.6	25.9	25.3	25.3	26.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	7	29.6	31.2	28.5	29.5	25.5	21.7	28.1	23.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	8	36.2	24.5	34.6	22.4	37.3	34.2	35.9	37.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	11	18.5	16.0	19.8	16.7	36.9	32.2	36.4	32.1	1.7	4.2	3.2	4.0	0.0	0.0	0.0	0.6
α = 0.10	3	48.8	32.4	51.4	33.1	41.1	41.7	41.3	40.7	1.0	1.0	1.1	1.4	0.0	0.0	0.5	1.1
	4	46.8	40.8	48.4	42.2	36.1	39.6	37.3	40.7	0.5	0.5	1.1	0.4	0.0	0.0	1.0	1.0
	5	30.8	30.5	33.8	32.1	42.6	37.9	42.5	38.3	0.5	0.5	1.1	0.4	1.2	1.2	1.0	1.6
	6	40.6	51.8	40.0	50.8	39.2	34.9	36.7	34.9	0.0	0.5	0.0	0.4	0.6	0.0	0.0	0.0
	7	40.9	38.7	38.2	37.7	39.8	34.2	41.3	37.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	8	46.8	39.4	43.9	36.6	48.4	46.0	50.3	47.9	2.1	0.5	2.0	0.5	0.6	0.0	0.6	0.0
	11	29.4	24.4	31.0	24.6	49.0	45.0	48.1	45.1	7.6	7.6	7.9	9.5	0.0	0.0	0.6	0.0
α = 0.20	3	53.1	36.2	56.0	38.0	45.8	45.2	45.0	45.5	2.4	1.9	2.8	2.5	3.0	3.0	5.3	4.8
	4	49.3	43.2	51.6	47.9	39.6	41.4	39.9	42.8	3.9	4.4	3.9	3.9	4.1	4.7	4.6	4.7
	5	35.9	33.5	38.5	35.4	45.0	41.4	44.6	42.5	5.1	4.1	5.4	4.4	4.7	4.1	5.2	4.1
	6	46.2	54.8	45.2	55.2	45.8	40.4	42.8	41.0	2.5	3.6	2.4	3.6	2.4	3.0	3.0	4.2
	7	46.8	41.9	43.5	40.1	43.5	37.9	46.1	39.5	2.2	1.6	1.9	1.4	3.7	2.5	3.0	2.4
	8	49.5	41.5	45.9	38.5	51.6	49.7	53.9	52.1	4.3	2.7	3.9	2.4	3.7	1.9	3.0	1.8
	11	41.2	37.8	42.1	38.1	53.0	48.3	51.9	48.1	12.6	15.1	13.5	16.7	4.7	4.0	4.9	3.7

There were vastly more schools and districts that experienced a decrease in the proportion of students with valid scores than an increase. If the decreases in specific schools had been due to chance, decreases and increases would be equally likely. Therefore, it is clear from these results that the proportion of students receiving valid test scores was affected by the logout interruptions, unless something other event occurred in both states simultaneously to depress the rate of students receiving valid scores. This seems highly unlikely. Therefore, it is highly probable that a considerable number of districts and schools have spring 2015 data from a sample of students that is likely not representative of the district or school.

The magnitude of decreases in schools and districts that experienced statistically significant decreases in the proportion of students receiving valid test scores are presented in Figure 11 (for schools) and 12 (for

districts). For these graphics we used a liberal statistical significance criterion, $\alpha = 0.2$, to capture as many likely decreases as possible. Schools without significant decreases were not included in these figures. In each of these figures (11 and 12), the decreases in percent of students receiving valid scores were put into 25 bins covering four percent (e.g., 0-4, 4-8, …, 92-96, 96-100). The dots on the plots indicate for the various grade levels the percent of schools or districts in the state that had a decrease that fell into each bin. A fourth-degree polynomial regression line was fit for each grade level in each panel to show the patterns in the distributions of the magnitude of statistically significant decreases in the percent of students receiving valid test scores[15].

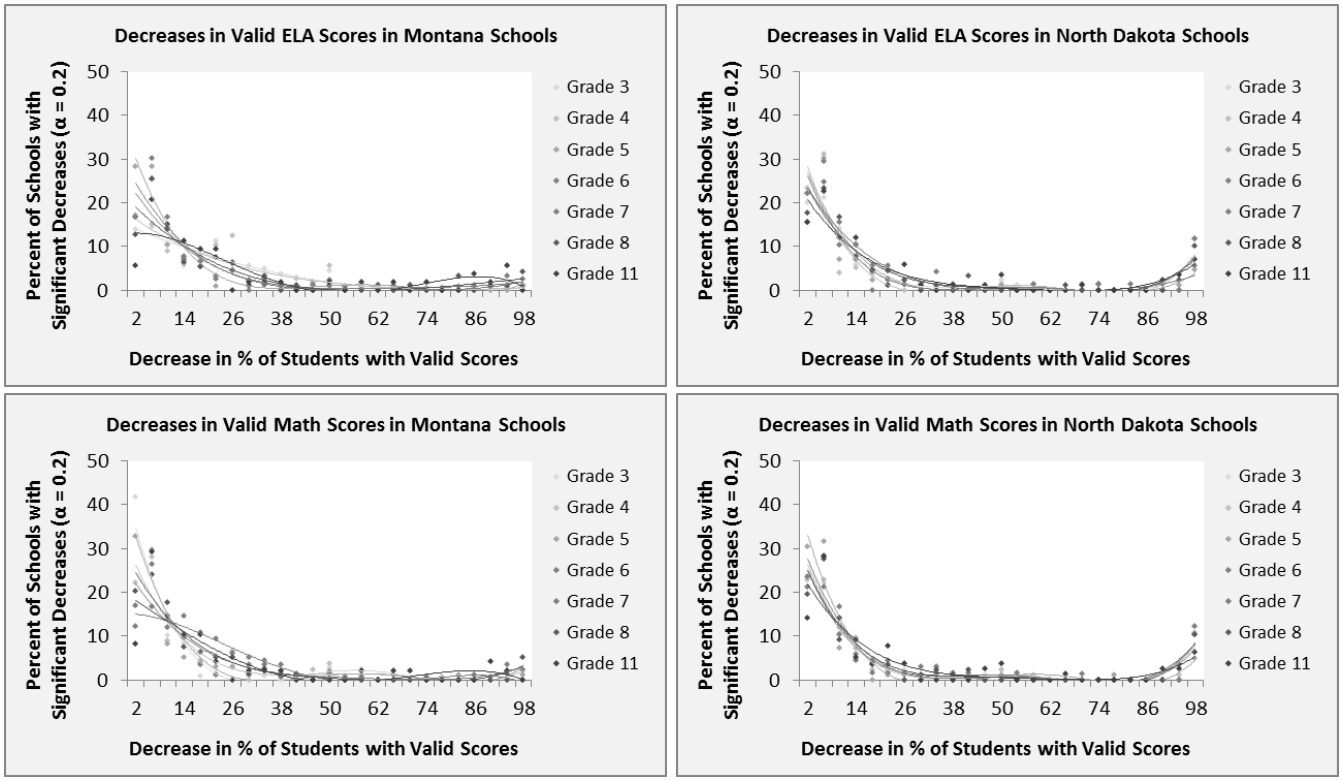

Figure 11. Magnitude of statistically significant *school*-level decreases in percent of students receiving scores.

As can be seen by comparing Figures 11 and 12, the distributions of magnitude of decreases in percent of students receiving scores were similar across schools and districts. In addition, the vast majority of statistically significant decreases in both school-level and district-level percent of students receiving scores were lower than 25 percent. However, there were a considerable number of districts and schools in both states with decreases in the percent of students receiving valid test scores of greater than 90 percent.

Finally, to clarify the typical effect of interruptions students receiving valid test scores, the average decrease in the percent of students receiving valid test scores (from the prior three-year baseline to 2015) is displayed in Table 12. Table 12 shows by state, subject, and grade the average decrease for both schools and districts. It also shows the average decrease for all schools/districts and for only schools/districts that had statistically significant decreases. As can be seen in Table 12, for all schools, the average decrease ranged from:

- 5 to 13 percent in ELA in Montana,
- 5 to 11 percent in math in Montana,
- 10 to 15 percent in ELA in North Dakota, and

[15] A 4th degree polynomial regression was used because it reasonably fit all data without overfitting.

- 15 to 21 percent in math in North Dakota.

For all districts, the average decrease ranged from

- 6 to 15 percent in ELA in Montana,
- 7 to 12 percent in math in Montana,
- 10 to 15 percent in ELA in North Dakota, and
- 11 to 16 percent in math in North Dakota.

For schools that experienced statistically significant decreases, the average decrease ranged from

- 19 to 48 percent in ELA in Montana,
- 22 to 41 percent in math in Montana,
- 34 to 52 percent in ELA in North Dakota, and
- 35 to 49 percent in math in North Dakota.

For districts that experienced statistically significant decreases, the average decrease ranged from

- 24 to 51 percent in ELA in Montana,
- 24 to 44 percent in math in Montana,
- 34 to 48 percent in ELA in North Dakota, and
- 35 to 43 percent in math in North Dakota.

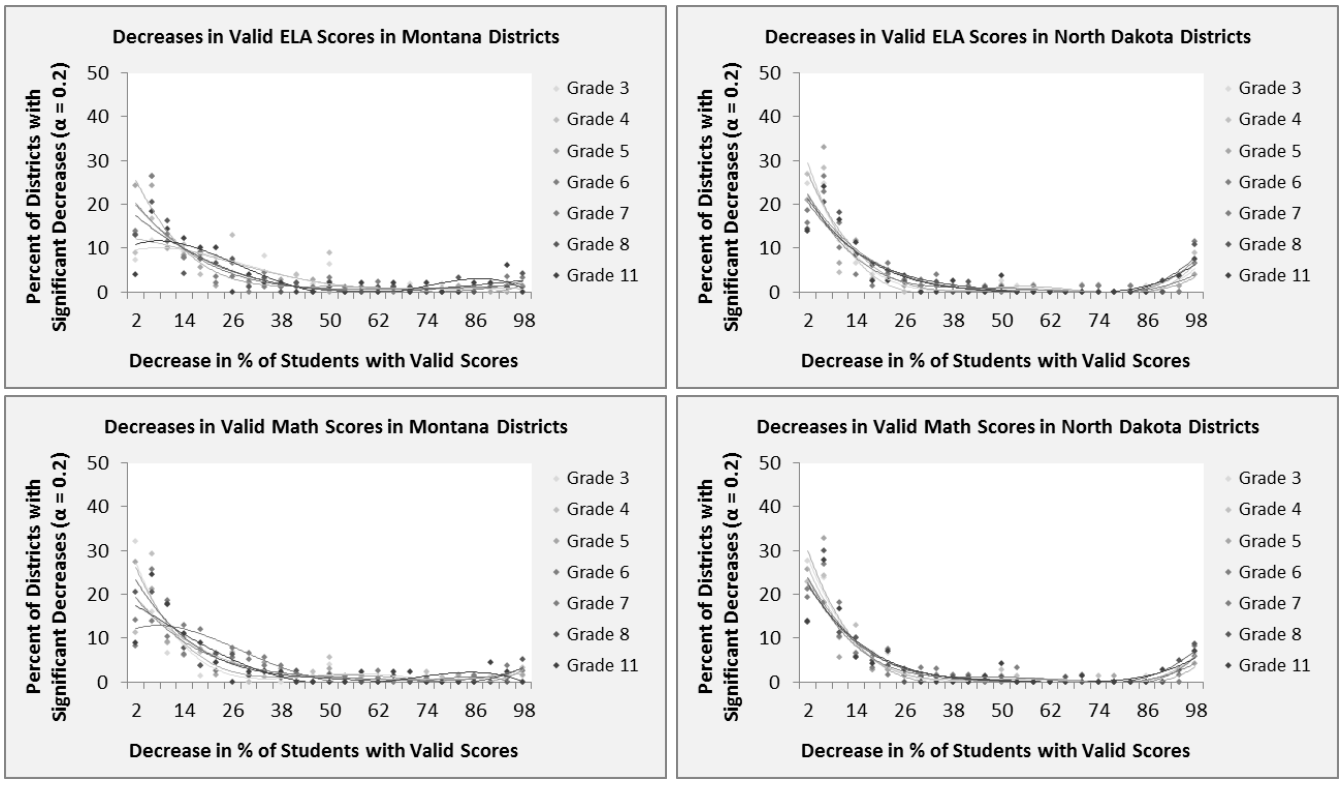

Figure 12. Magnitude of statistically significant *district*-level decreases in percent of students receiving scores.

The average decreases for all schools/districts in both states appear to be consequential, while the average decreases for schools/districts experiencing statistically significant decreases in both states appear to be substantial. States will need to determine whether the spring 2015 deviations from the baseline rates of students receiving valid scores are large enough to warrant not using the aggregate school and/or district results for high stakes purposes. To aid states in making this determination, the increase/decrease determinations using the three confidence levels are provided.

However, these decreases in participation are not as great as they could have been. Schools were granted flexibility very early on in the administration windows – before most students tested. If the schools in each state had exercised this flexibility by making a blanket decision not to test, then the decreases in participation rates would have certainly been greater than those found here (in Montana, there was a decrease in participation of 10-11% in ELA and 7-9% in mathematics; in North Dakota 12-13% in ELA, 13-17% in mathematics). Thus schools appeared to be judicious in their use of flexibility in order to keep participation rates as high as possible.

Table 12. *Average Decreases in Percent of Students with Valid Scores from the Prior Three-Year Baseline.*

Subject	Grade	Average Decrease in Percent of Students with Valid Scores for…							
		All Decreases				Statistically Significant Decreases			
		Montana		North Dakota		Montana		North Dakota	
		Schools	Districts	Schools	Districts	Schools	Districts	Schools	Districts
ELA	3	13%	15%	15%	11%	27%	33%	51%	38%
	4	10%	12%	13%	10%	25%	29%	52%	42%
	5	5%	6%	12%	10%	19%	24%	43%	34%
	6	8%	10%	10%	11%	25%	32%	34%	37%
	7	9%	11%	14%	13%	27%	30%	45%	48%
	8	11%	13%	14%	14%	30%	33%	36%	35%
	11	10%	10%	15%	15%	48%	51%	38%	38%
	All	***10%***	***11%***	***13%***	***12%***	***27%***	***32%***	***43%***	***39%***
Math	3	5%	8%	21%	13%	25%	30%	49%	35%
	4	8%	9%	18%	12%	24%	27%	45%	37%
	5	6%	7%	15%	11%	25%	30%	43%	35%
	6	11%	12%	16%	14%	22%	24%	36%	36%
	7	9%	10%	17%	14%	27%	31%	47%	43%
	8	6%	7%	19%	16%	24%	27%	36%	35%
	11	7%	8%	15%	12%	41%	44%	35%	35%
	All	***7%***	***9%***	***17%***	***13%***	***25%***	***29%***	***41%***	***36%***

RECOMMENDATIONS FOR AMELIORATING DATA LIMITATIONS

As noted previously, limitations in the collected data as well as the lack of documentation reduced the scope of this work considerably. However, these issues are symptoms of a border problem - treating interruptions of online testing as an *abnormality* and responding in an *ad hoc* manner *if* it happens.

Build Systems to Treat Interruptions as Inevitable, with Mechanisms to Respond Quickly, Efficiently, and Effectively

Because of the high degree of complexity and dependence on external systems (such as internet service providers, routers, district infrastructure), it does not makes sense to treat interruptions as abnormalities. They should be treated, rather, as *inevitabilities* and vendors and states should plan *in advance* how to respond effectively and efficiently *when* interruptions happen. An important aspect of such a plan is to have data systems that represent not just the transactions between client and server systems but that also include specific data elements to allow for a reconstruction of what students actually experienced while testing,

instead of just data elements that explain operations indirectly related to student experience (e.g., when a set of items was downloaded into the memory of the application administering the test to the students).

Because online assessment will always come with the risk of both system-wide and idiosyncratic interruptions, it is imperative that software and data systems be updated to implement the following critical improvements:

1. Real-time notification regarding various types of system issues that could lead to interruptions available to state, vendor, and local school staff.
2. Storage of data elements that appear to be missing from the current versions of software and data systems, which allow for the types of analyses that were originally proposed in this work.
3. Transparent documentation of all aspects of software and data storage that is readily understood by qualified information technology professionals.
4. Rapid extraction of data so that analyses can be performed quickly enough to inform states about appropriate actions before reporting assessment scores. A preferable alternative is to build analysis functionality directly into the system, to aid states in understanding the prevalence and effects of interruptions without the need for specialized data extraction and data analyses.

Category 3 is a universal recommendation critical to improving states' ability to respond to test interruptions. Specific recommendations that address the concerns captured in categories 1, 2, and 4 are provided in the next three sections. The first section describes specific data elements that need to be captured to support the investigation of interruptions and to improve the ability to respond to interruptions. The second describes how interruptions can be taken into account as part of the design of the testing system, under the auspices of disaster recovery. The third section describes how a system designed with interruptions should interact with users when interruptions actually do occur. Finally, the specific data elements that any test administration system should include in order to support investigation of interruptions and appropriate rapid response is provided in Appendix A.

Expand the Definition of Disaster Recovery

In large-scale testing, disaster recovery is typically understood as successfully stopping a critical problem and then successfully completing the rest of testing. In the context of interruptions, such a definition is insufficient. It is insufficient because it leaves unaddressed the effects of interruptions on student test scores, test completion, and test participation. Those effects cannot be appropriately mitigated unless states and vendors can immediately and appropriately respond to interruption events. One reason that interruptions have potentially dire implications for testing programs is that states do not, reasonably, have the ability to immediately and appropriately respond to interruptions. This is largely because current systems have not been developed to facilitate such responses. However, this is also because both states and their vendors have treated interruptions as *abnormalities* rather than *eventualities*.

This means that in addition to what we recommend below, it is important that the data collection systems and the score reporting systems be enhanced.

The data collection systems must be able to document each type of issue that a vendor and its clients can reasonably anticipate. Steps to insuring the data collection system can do so include:

- For each type of issue that can be reasonably anticipated, the vendor and its clients must identify the data elements that need to be stored to adequately analyze and respond to the issue. Steps must then be taken to insure that these data elements are recorded by the system during administration.

- Making data structures flexible enough that novel issues (e.g., unanticipated types of interruptions) can be flagged and documented when they occur, rather than having to write new queries or rewrite existing queries to produce data on novel issues. This should include several data fields that can be used flexibly to document nuances for each type of novel issue (e.g., a notes or comments field that can be used to store contact information for a person who reported an issue, explanations of nuances).
- Data file and report generation should be enhanced, so that data files and reports are flagged when issues like interruptions occur. These flags should be associated with text that explains what issue occurred and what implications that issue may have for test score interpretations.

In addition post-administration scoring and reporting systems must also be improved to accommodate issues as they occur:

- Make scoring and reporting systems flexible enough to hold individual items out associated with a given issue from scoring without delaying scoring for the remainder of a student's test or the scoring of other students' tests.
- Make scoring and reporting systems flexible enough to adjust, as necessary, a score and its associated measure of confidence for the effects of interruptions (or other issues in administration).

Increase Information Available about Interruptions at All Levels

When responses to interruptions in online testing are not planned in advance, confusion can be expected. State and vendor staff are already under considerable pressure when students are participating in testing. Without advance planning, responses to interruptions are likely to be ad hoc and less than optimal because both state and vendor resources are stretched thin while students are testing. Existing systems tend to leave school, district, regional and state staff, as well as the students themselves, in the dark when an interruption occurs. It is also likely that many vendor staff are themselves in the dark when interruptions occur. This lack of information increases the anxiety of everyone involved in the testing process and may lead to sub-optimal decision making.

To address this issue, as a part of redefining disaster recovery, we additionally recommend that systems be modified to include the following features:

- Clear and thorough documentation of all data tables and data elements including, for applicable data tables and elements, a description of their connection to student experience interacting with the test administration interface and the device used for test administration.
- Server applications capable of detecting and sharing real-time information about interruptions at all levels (e.g., vendor, state, regional school staff, district staff, school staff, students).
- Applications that do the same thing as the previously mentioned server applications, but on the client side. That is, client applications capable of detecting and sharing real-time information about interruptions on each student's device with server applications. This redundancy insures that if a server application or a server database fails, the data remain captured on student devices. In terms of the current project, the server applications became overloaded and stopped writing data into the databases. Thus no data was captured during system-wide interruption events. Client side applications capturing data locally would have helped ameliorate the problems that ensue from this type of failure.
- Independently operating databases on independent servers that exist for the sole purpose of documenting issues with test administration. These databases and servers should be independent of the databases and servers that perform the work of test administration. These independent databases

should collect information from server applications, server databases, and client devices to document administration issues like interruptions.

- Applications installed on devices used by key vendor staff, key state staff, key district and school staff (e.g., test coordinator, proctor, administrator, etc.) should display real-time information about issues with test administration for all students for whom the user is responsible. The data should be gathered from the independent databases on the independent servers described above.
- When issues are detected, the real-time reports on devices used by vendor staff, state staff, and district/school staff should include the following at a minimum:
    - What occurred (including whether it is known what occurred)
    - The implications for the student (such as indicating whether the student can pick up where s/he left off when the system is available again; the degree to which data has been lost if any)
    - The prevalence of the issue
    - Instructions for next steps
    - Contact information for obtaining further information
- When issues are detected, the real-time reports on devices used by students should include the following at a minimum:
    - What occurred (including whether it is known)
    - What the implications are for the student (including whether it is known)
    - Instructions for next steps

REFERENCES

Bynum, B.H., Hoffman, G., Thacker, A., & Swain, M. (2014). *A statistical investigation of the effects of computer disruptions on student and school scores*. Presentation at the Council of Chief State School Officers (CCSSO) annual National Conference on Student Assessment, New Orleans, LA.

Dodge, Y. (2010). Mahalanobis Distance. In *The Concise Encyclopedia of Statistics* (pp. 325-326). NY: Springer-Verlag.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Kim, J. & Seltzer, M. (2007). *Causal inference in multilevel settings in which selection processes vary across schools*. (CRESST Report 708). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Martineau, J., Domaleski, C., Egan, K., Patelis, T., & Dadey, N. (2015, November). *Recommendations for addressing the impact of test administration interruptions and irregularities*. Washington, DC: Council of Chief State School Officers (CCSSO). Available online: http://www.ccsso.org/Resources/Publications/Recommendations_for_Addressing_the_Impact_of_Test_Administration_Interruptions_and_Irregularities.html.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Hong, G & Raudenbush, S. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis,* 27(3), 205-224.

Rubin, D. (2006). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20-36.

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2011). *HLM 7: Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Sinharay, S., Wan, P., Whitaker, M., Kim, D.-I., Zhang, L., & Choi, S. W . (2014). Determining the overall impact of interruptions during online testing. *Journal of Educational Measurement*, 51(4), 419-440.

Sinharay, S., Wan, P., Choi, S.W., & Kim, D.-I. (2015). Assessing individual-level impact of interruptions during online testing. *Journal of Educational Measurement*, 52(1), 80–105.

Appendix A: Specific Data Elements Needed to Support Investigation and Appropriate Response

Rather than requiring supplemental database queries and subsequent analyses of the resulting data, the data elements listed below should be stored directly for each student. In addition, this stored data should be easily accessible via a prewritten "canned" query.

- Student ID (and associated school, district, and demographics)
- Each time a student logs onto the system (i.e., for each unique log in)
    - Timestamp for signing on to the system (e.g., the time the student submits login credentials).
    - Timestamp for signing out of the system.
    - Type of logout (involuntary, voluntary)
    - Cause of involuntary logout (including unknown).
- Timestamp for when the student was presented with the page(s) containing instructions and/or other test matter.
- Timestamp for when the student clicked the button/screen element to take him or her to the first item on the test.
- Timestamp for each click of the button/screen element to take the student to the next page of items
- For each item on the student's test
    - Item ID
    - Student score on the item
    - Item number (i.e., the order in which the student saw the item)
    - For each unique time the student visited the page on which the item was displayed
        - Timestamp for when the page was displayed (good) or timestamp for when the content of the item itself was displayed (better)
        - Time stamp for when the student first interacted with the item
        - Time stamp for when the student last interacted with the item
        - Time stamp for when the student navigated away from the page on which the item was displayed
    - Student's provisional score after the item was scored.
- Timestamp for when the student submitted the completed test.
- IDs of all the items administered, in order, including navigating back and forth across pages of the test (meaning that duplicate item IDs are acceptable).
- For each interruption experienced
    - The type of interruption
    - The time of the interruption
    - The ID numbers of the items already answered.
    - The student's provisional score before the interruption.