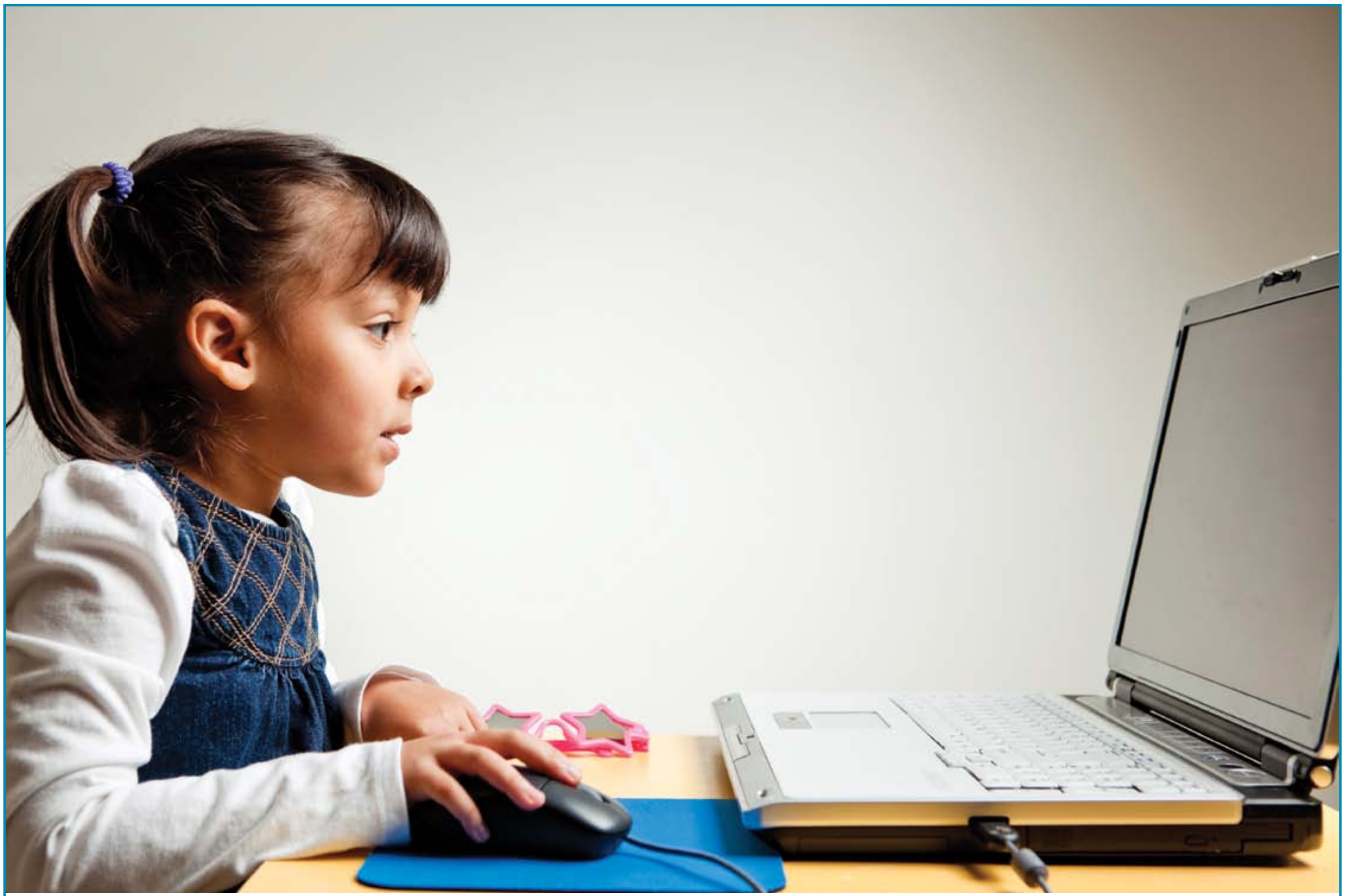


# Recommendations for Addressing the Impact of Test Administration Interruptions and Irregularities

CCSSO
Council of Chief State School Officers

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

*Recommendations for Addressing the Impact of Test Administration Interruptions and Irregularities*

**National Center for the Improvement of Educational Assessment**
Joseph Martineau
Chris Domaleski
Karla Egan
Thanos Patelis
Nathan Dadey

*We thank Richard Hill for his comments and contributions to this paper and the body of work in this area.*

**COUNCIL OF CHIEF STATE SCHOOL OFFICERS**
June Atkinson (North Carolina), President
Chris Minnich, Executive Director

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

# Contents

# Introduction

Computer-based assessments are not new, having been used by some professional and licensure exams for over 20 years. In the world of K-12 education, administering computer-based summative assessments is rapidly becoming the norm instead of the exception. This move has presented new opportunities and challenges for the states and vendors administering these computer-based assessments. While much has been written about technology-enhanced items or computer-adaptive testing, less attention has been paid to the challenges that arise because of technological glitches. Chief among these challenges is the testing interruptions that occur when technology fails.

Over the past few years multiple states and vendors have experienced technological glitches during the administration of computer-based assessments. In 2015, testing interruptions occurred in Montana, Nevada, North Dakota, and Georgia. Between 2010 and 2014, similar interruptions occurred during the online administration of statewide assessments in Florida, Indiana, Kentucky, Minnesota, Oklahoma, and Wyoming. While there are various ways for both small- and large-scale test interruptions to occur with paper and pencil testing (e.g., school-wide fire drill, statewide natural disaster), computer-based testing adds complexity and creates more opportunities for test interruptions, particularly large-scale interruptions.

When interruptions occur, the primary concern is the impact on (a) student performance and (b) student participation. In terms of the former, student test scores are usually the unit of analysis, although we also suggest follow up item level investigations. A student's test score may be affected if, for example, she is kicked-off of the administration platform or stuck on a single item while the platform stalls. However, the nature of the impact on student scores may not be straightforward. A student's performance may be disadvantaged, perhaps due to a break in concentration or general frustration. Or, performance may be advantaged by giving the student additional time to think about the content or, even, look up answers to items while the test administration system is offline or suffering a slowdown. These types of issues extend to the latter area of concern — student participation. A student who has been interrupted may become so frustrated she may refuse to finish the test. A teacher, school, or district may decide not to have their students participate in the assessment. Or, the state may invalidate the test following administration.

The purpose of this paper is to provide a general framework for examining the impact of technology-related interruptions on student scores on statewide summative assessments. As such this paper is meant to act as a guide to structuring investigations of interruptions and thus does not advocate for the use of any specific statistical model. We first define and provide examples of what is meant by a technology-related interruption. Next, we review the current literature and indicate the methodologies and data used to evaluate the impact of interruptions. Then the main portion of the paper provides a framework for examining how technology-related interruptions may have impacted student scores. Finally, we address implications and alternatives for reporting and accountability based on the outcome of the analyses.

## Defining Technology-Related Interruptions

Before examining the possible impact that interruptions have on student scores, it is important to define what is meant by technology-related interruption. Simply stated, a technology-related interruption is an event that disrupts students' testing experiences caused by the computers, online systems, or other technological devices through which the test is delivered. Table 1 lists and provides examples of some common types of technology-related interruptions.

*Table 1. Common Types of Technology-Related Interruptions*

| Type of Interruption | Example |
|---|---|
| **Delayed log-in** | The student is unable to access the system at the scheduled time to begin or resume testing. |
| **System delays in loading items** | After submitting a response to an item, the system 'freezes' for an unusually long period of time (e.g., 30-60 seconds or more) before advancing to the next item. |
| **System slows down** | System lags behind student input often resulting in jerky and delayed response to student commands. |
| **Students are unexpectedly logged out** | During test administration, the student is unexpectedly 'exited' from the system and must log-in again to resume testing. |
| **Students lose all progress** | During test administration the system fails and the student loses all of his or her progress, resulting in that student retaking a portion of the test (or not having any score information recorded). |
| **Item(s) or stimuli is (are) rendered poorly** | An image is incorrectly resized on some devices, causing it to be distorted. |
| **Resources not available or not working properly** | The highlighter tool is not available for all or a portion of the administration time. |
| **Accessibility feature(s) not working properly** | The read-aloud feature is not functioning for a portion of the administration time. |

Importantly, these examples do not reflect the full range of 'glitches' that can occur with technology. We focus here on technology-related failures that impede or delay interaction with test content, but we do not include failures that fully prohibit interaction or that introduce an unquestionable error. For example, we do not list problems such as an item without a correct response. While we include items or stimuli that render poorly, we do not include missing content. Our intent is to describe conditions that may impact the test taking experience in a way that is not intended. Such conditions may be thought to frustrate the examinee or make it more challenging for the student to demonstrate their knowledge and skills. However, the direction and magnitude of the impact are unknown, hence the need for the additional investigation described in this document.

It is important to realize that there are more things in a computer-based testing program that can lead to interruptions relative to a paper-based testing program. We advise states to develop contingency plans to avoid or minimize interruptions (e.g., a redundant system immediately starts when a connection is lost). Even with minimal interruptions, it is the responsibility of the test user (in this case the state testing agency and its vendor) to document any disruptions to the standardized test administration procedures (see Standard 6.3 and 6.9 in AERA, APA, & NCME, 2014) and to ensure that the disruptions do not compromise the inferences made based on the reported scores.

## Previous Studies of Test Interruptions

As indicated in the previous section, there are a number of types of interruptions, each of which could have a broad range of potential effects on student scores. In hindsight, the causes of interruptions can seem obvious, but test interruptions are hard to predict and can occur at any time. There may be some situations that can increase the likelihood of disruptions in the testing environment (e.g., electrical storm, older computer connections, new and untested applications), however, most are difficult to predict and prevent. Additionally, even if there are redundant systems and contingencies that permit the seamless transition of the administration of the test, because the test scores have important uses, examining the potential impact of the test interruptions (even with an immediate contingency occurring) becomes an important if not compulsory undertaking.

### DATA

Because the issues that cause interruptions are difficult to predict, the entity responsible for the implementation of the testing program must collect data about the student and the testing environment. This way, interruptions can be identified and documented. Additionally, student progress during the test must be closely monitored, so that the interruptions can be identified quickly and a decision can be made as to whether students should be allowed to continue.

Before we review the methods used in previous situations where test interruptions have occurred, we believe it is important to mention the data elements needed for these methods. The key elements used by prior studies include identifiers for students and schools; indicators of interruptions; assessment scores from the prior year(s); and the current assessment scores. Some methods also use student item responses, an indicator of when during the test the first interruption occurred and the current assessment's pre-equated item parameters. Studies that use matching methods, like propensity score matching also use student demographics. In addition, we also suggest additional data elements that are needed to support the methodological approach we suggest. The extent to which data are available will permit the methodologies to be fully implemented and greatly increase the number of strategies that can be used to evaluate the impact of the test interruptions. Limitations in the data elements available will also limit the methods that can be used.[1]

[1] Based on our past experience working with large-scale data systems, care needs be taken not only to collect and provision the data, but also to insure the data are organized and formatted in a manner that can easily be ingested by the analytical applications used. If the data format is not considered and designed ahead of time, significant time and expense usually occurs in scrubbing, formatting, and organizing the data for use in the analysis applications and related software.

The general categories of the data along with a brief description are shown in Table 2.

*Table 2. Type of Data Elements Needed to Examine Test-Based Interruptions*

| Source | Type | Description |
|---|---|---|
| Student | ID | A student identifier that can be linked to other sources of information about the student (including previous years' test data). |
| | Demographic | Information that identifies student as a member of relevant subgroups including (a) gender*, (b) race/ethnicity*, (c) English language learner status*, (d) socio-economic indicator* (e.g., free/reduced lunch indicator), and (e) special education category. |
| | Interruption | Data fields that define the interruption (a) type/severity, (b) start date/time*, and (c) end date/time* experienced at less than a system-wide level (could be derived from a school or district-level file if the interruption affected an entire school or district). |
| | School and District IDs | The school and district identifiers associated with the location of the student's primary learning environment. Type of school*. |
| | Experiential | Any educationally-relevant information should be captured. This information will need to be associated with a date indicating the academic year. This includes grade level, courses taken, academic performance (e.g., grades or grade point average), current test scores in other subjects*, and test scores from previous years*. |
| Test | Student ID | Student identifier that links the item and assessment data to other student records (like those containing demographic data). |
| | Item Metadata | Data associated with the items including (a) item ID*, (b) date and time administered*, (c) time of answer submission*, (d) item content area/standard, (e) pre-equated item parameters*, (f) testing session, (g) section, (h) position, (i) scoring, and (j) code for scored vs. experimental. |
| | Item responses | For each student, unscored item responses including codes for no response. For each item, indicate whether the item response changed from wrong to right. |
| | Item scores | For each student, scored final item responses including codes for no response*. For each student, provide the number of items changed from wrong to right*. |
| | Scores | Student raw score, scaled score*, performance level, subscores and section scores (for multistage/section tests). |
| System Interruption | Interruption | Data fields that define the interruption (a) type/severity, (b) start date/time, and (c) end date/time for system wide interruptions. |

** Indicates the variable has been used in prior studies of interruptions.*

## METHODS

Previous studies have generally addressed the following question, "What score would students, who experienced an interruption, have gotten had not been interrupted?" Hill (2010, 2013a, 2013b) addressed this question through a number of comparisons to historical data, particularly prior test scores. Sinharay et al. (2014, 2015) and Bynum et al. (2014) address this question by using several different models to predict scores for interrupted students, generally by using student item responses before an interruption. These comparisons were done at the student and aggregate levels (e.g., school and state).

The methods used and summary of the results for each of these published studies are presented in Table A-1. While various methods were used, they all used student-level data collected during the test and previous administrations. All of the studies made comparisons between students experiencing interruptions and students not experiencing interruptions, by grade, and by content area. In three studies (Bynum et al., 2014; Sinharay et al., 2014, 2015), propensity score matching was used to select comparable students representing those not experiencing interruptions.

Four of these studies used the same state data (i.e., Hill, 2013a, 2013b; Sinharay et al., 2014, 2015) and all of the methods used came to the same conclusions regarding the size of the impact of the interruptions — the impact of interruptions was small to minimal.

# Methodological Recommendations

There are at least four issues with technology-based interruptions that are only partially addressed by previous methods.

First, even if the uninterrupted and interrupted samples are demographically equivalent (e.g., interruptions are truly randomly distributed across districts, schools, and students; or propensity score matching has been used), it remains important to include at least some demographics in subsequent analyses. This is because students of varying demographic characteristics may be affected differently by test interruptions. In one hypothetical example, a high-achieving student might take the time to investigate items on the section of the test she was currently taking, while a low achieving student might become very frustrated by the experience and not take the opportunity offered. Assuming that the effects of interruptions are the same for all demographic groups may make it less likely that real effects could be detected.[2] While most demographics related to test scores are reasonable to investigate, we suspect the most important to be those that could reasonably be expected to show differences in the levels of stress and anxiety produced by a test interruption, such as prior achievement level and characteristics such as disabilities or limited English proficiency that tend to reduce access to test content.

[2] Addressing this could be accomplished by performing descriptive analyses separately by group, or in a statistical model, by including interactions between interruptions and demographics.

Second, aggregate scores may be affected by interruptions even if interruptions are not predictive of any available individual or aggregate scores. Interruptions may have effects on test *participation/completion* rates that in turn inflate or deflate aggregate scores (e.g., some students are not tested or do not complete enough of the test to produce a score as a result of interruptions, and selection into this group of students without scores is related to level of achievement). Therefore, using only available test scores to investigate effects on aggregate scores is insufficient. The potential inflation or deflation of aggregate scores resulting from non-availability of scores attributable to interruptions should also be considered.

Third, various methods have important differences in the questions they are designed to answer. Some methods are designed to answer *for each individual student what his or her score would have been had no interruptions been experienced* (e.g., Sinharay et al., 2014, 2015). Other methods are designed to answer for each student what his or her score would have been had no interruptions been experienced *assuming that his or her response to interruptions was not idiosyncratic.*[3] The first question is a more useful one (as it allows for idiosyncrasies in response to interruptions), but requires a certain data structure that may not always be available.

Methods designed to answer the first question split students item score strings based on which items were unaffected (responded to before interruption) and those that were affected (responded to after interruption) and estimate pre-interruption scores (e.g., ) as a baseline for comparing against post interruption performance. This presents two important limitations. First, the degree to which such methods will perform well is a function of how stable the estimated pre-interruption score is. Therefore, these methods can be expected to work best for the students for whom interruption effects would be expected to be the smallest (i.e., students with a larger proportion of unaffected items unaffected.[4] In addition, the pre- and post-interruption dichotomy maybe untenable for computerized adaptive tests (CAT). Second, for this type of design to work, the item order of the interrupted students must be matched by at least one uninterrupted student (preferably many uninterrupted students to allow for matching on other important predictors such as prior test scores and demographics). This type of match is not guaranteed under a CAT design. Each interrupted student could, in theory, have a unique ordering of items and this ordering may not have a match from the uninterrupted student sample, let alone the ideal of several matches.

Fourth, the pre- and post-interruption dichotomy may be insufficient. It may be important to separate the effects of submitting an item during a test interruption (e.g., a system slowdown rather than a system failure) vs. after an interruption, in that the in-the-moment frustration of suffering a slowdown may be greater than after a slowdown has ended. In addition, there may be multiple types of interruptions or multiple interruptions of the same type. Methods capable of investigating effects beyond a single before/after interruption dichotomy may be needed.

Based on previous methods used and possible limitations of those methods, a framework for analyses of test interruptions is presented in Figure 1. For any given study, the analyst(s) will need to create codes to

[3] That is, all students with the same prior test score(s), demographic profile, and experience of interruption(s) respond to interruptions in the same manner.

[4] Assuming there are no true differences in student achievement of content presented pre- and post-interruption.

classify the various types of interruptions students in their data set encountered (Table 1 is by no means comprehensive, but may provide a start). Given the complications that many types of interruptions create for analyses, it will be important to consider what interruptions are important to investigate separately, and which types can be collapsed into combined categories.

In addition, it is important to make a distinction between a student directly experiencing an interruption (the student's own test event was subject to the interruption) and indirectly experiencing an interruption (the student did not directly experience the interruption, but a student in the same administrative unit did; see Hill, 2013a, p. 4). An administrative unit might be defined as a test session, class, school, or district. Effects of direct and indirect interruptions of the same type should be investigated separately to account for direct interruptions being more likely to affect scores than an indirect interruption. For investigating effects of indirect interruptions, it is reasonable to expect that investigations will be more sensitive the smaller the definition of the administrative unit (where it is more likely that other students will become aware of another student's direct interruption).

Based on this delineation of types of interruption, when we use the phrase "type of interruption" we mean all of the following:

- Directly experiencing each specific type of interruption
- Directly experiencing at least one type of interruption
- Indirectly experiencing each specific type of interruption
- Indirectly experiencing at least one type of interruption

The framework in Figure 1 was developed to answer the following questions:

A. How prevalent were the various types of interruptions for different important groupings of students? (Step 2)

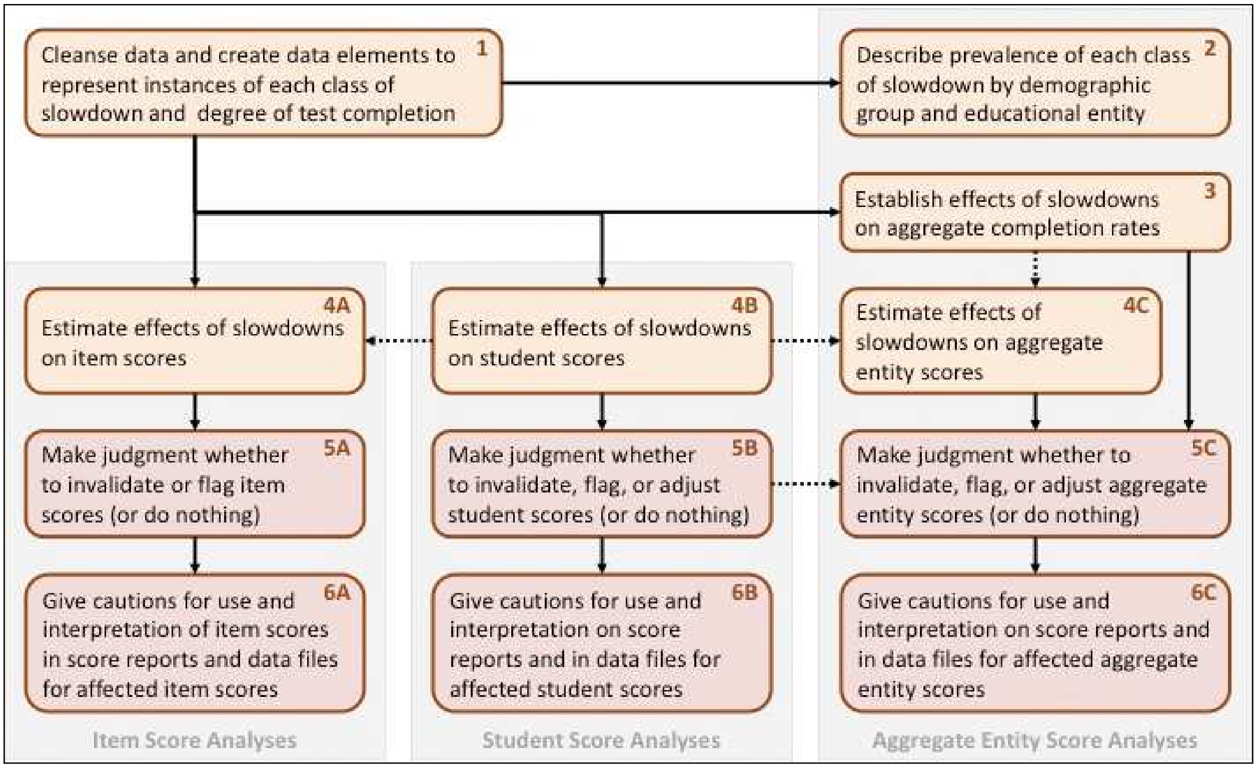
B. Did slowdowns affect the rate of test completion, and do any effects on test completion inflate or deflate aggregate scores? (Step 3)

C. How were item, student, and/or aggregate scores for tested students affected by interruptions? (Steps 4A-4C)

D. How should item, student, and aggregate scores subject to various types of interruptions be treated in reports and data files? (Steps 5A-5C)

E. What guidance regarding interpretation and use should be provided regarding item, student, and aggregate scores subject to any interruption and to interruptions found to have effects on scores? (Steps 5A-6C)

*Figure 1. Framework for Analyzing Effects of Test Interruptions and Determining Appropriate Treatment of Potentially Affected Scores*


Cleanse data and create data elements to represent instances of each class of slowdown and degree of test completion 1
Describe prevalence of each class of slowdown by demographic group and educational entity 2
Establish effects of slowdowns on aggregate completion rates 3
Estimate effects of slowdowns on item scores 4A
Estimate effects of slowdowns on student scores 4B
Estimate effects of slowdowns on aggregate entity scores 4C
Make judgment whether to invalidate or flag item scores (or do nothing) 5A
Make judgment whether to invalidate, flag, or adjust student scores (or do nothing) 5B
Make judgment whether to invalidate, flag, or adjust aggregate entity scores (or do nothing) 5C
Give cautions for use and interpretation of item scores in score reports and data files for affected item scores 6A
Give cautions for use and interpretation on score reports and in data files for affected student scores 6B
Give cautions for use and interpretation on score reports and in data files for affected aggregate entity scores 6C
Item Score Analyses
Student Score Analyses
Aggregate Entity Score Analyses

Though judgment calls are required in designing analyses, steps 1-4C can be conducted with some degree of objectivity. Regardless of whether practically important effects of interruptions are found in steps 3-4C, policy makers (with the support of assessment experts) must make policy judgments in steps 5A-6C (captured in the boxes shaded in light red in Figure 1) informed by the results of prior steps. Finally, in Figure 1, some steps may not be needed, depending on the claims one intends to make based on the results of the analyses. For example, several studies (e.g., Bynum et al., 2014; Hill, 2010, 2013a, 2013b; Sinharay, 2014) only examined effects of interruptions at the aggregate level.

Following Figure 1, the purpose of each step (numbered in the top right corner of each step) is more completely described. We purposely avoid detailed recommendations for each step to keep the paper focused on a framework. Any descriptions of study designs are general to allow for analysts to create designs best suited to a specific context.

## PRELIMINARY STEPS

**Step 1** is intended to verify that data are clean and to create the necessary additional data elements to conduct analyses. New data elements generally include the following variables:

*Student-level variables*

- The degree to which each student completed the test
- Whether each student experienced each type of interruption
- For each type of interruption, a variable indicating the proportion of completed items that were submitted before, during, and after each type of interruption.[5],[6]

*Student- by Item-level variables*

- Whether each student submitted each item
- For each type of interruption, a variable indicating whether each submitted item was submitted by each student before, during, or after interruption

Additional new student- and aggregate-level variables will also be created in later stages of the analyses.

**Step 2** is intended to determine the prevalence of each type of test interruption for student subgroups (e.g., gender or English Language Learner status) and educational entities (e.g., school or district). While the statistics are produced primarily for transparency in reporting, they may also be useful to identify subgroups or entities which should be closely examined in steps 4C, 5C, and 6C.

## TEST COMPLETION ANALYSES

While it is impossible to definitively establish whether interruptions caused a change in participation/completion rates for any individual school or district, it is important to examine whether interruptions are related to test completion and to identify schools, districts, and/or subgroups for which it is reasonable to suspect that completion rates were affected by interruptions.

We present two broad approaches for doing so. The first is an analytical approach in which the relationship of current year completion/participation with prevalence of direct and indirect interruptions is investigated relative to historical participation rates and the historical relationships between participation rates and group demographics. This is a more rigorous approach to the problem.

[5] In some instances, it may not be possible to make the before/during/after interruption distinction. For example, when a student is logged off involuntarily, the only possible distinction may be before/after interruption. Multiple interruptions of the same type complicate data creation. In such cases the three-category classifications may be created to indicate before the first interruption, during an interruption, and after the first interruption has ended. Two-group classifications may indicate before the first interruption and after the first interruption. Another approach is to note the number of interruptions experienced before submitting an item (0 to N), and whether an interruption was underway when the item was submitted. Three-category classifications can also complicate propensity score matching, as multiple levels of and/or multiple "treatment" effects can be difficult to manage.

[6] Requires data specifying the date and time at which each item is presented to the student, as well as the date and time of the beginning and ending of each interruption. These times are compared against each other to create the necessary flags.

The second is a heuristic approach in which current-year completion/participation rates are compared only with historical statewide and/or same-group participation rates (e.g., schools, districts, subgroups) that had a sizable proportion of students experiencing slowdowns. In this case, groups could be flagged as affected by examining whether current completion/participation rates are statistically significantly lower than mean historical same-group and/or state participation rates. This approach better matches the more heuristic approach to participation rates under federal law (the arbitrary selection of 95% as the target participation rate). As noted in Figure 1, step 3 influences the estimates of slowdowns on aggregate scores.

## ITEM, STUDENT, AND AGGREGATE SCORE ANALYSES

It may be important to investigate effects of slowdowns on all three levels of scores. We start with analyses of interruption effects on student scores (**Step 4B**) because (1) they are the primary scores of interest, and (2) identifying effects on student scores can considerably simplify the analyses of effects on item scores and on aggregate scores.

While we present the student- and aggregate-levels separately, the two are related. Most prior studies have only examined slowdown effects at the aggregate level. However, if student level analyses are also conducted, the student level impact can be aggregated and used to inform the aggregate analyses (see Sinharay et al., 2015, p. 87). In addition, if both student- and aggregate-level scores are under investigation, policy makers may want to apply similar decisions about what to do with potentially affected student and aggregate scores. This eliminates the need to explain that different procedures were applied at student and aggregate levels. Finally, since the judgments made about the scores in steps 5B and 5C, as well as the cautions in 6B and 6C, are similar, we have grouped our discussion of them.

**Step 4B** is intended to identify effects of direct and indirect interruptions on each student's score. Most of the methods[7] for examining student-level impact require an indicator of when the first interruption occurred, and may also require item response data (see Sinharay et al., 2015). However, these methods that require item response data have yet to be extended to CAT and it is unclear if these methods *can* be extended to CAT. One possible way to address this issue is through analyzing different degrees to which student test scores are affected based on an estimated proportion of items that could reasonably have been affected by an interruption (those administered during/after an interruption).[8]

There are at least three analytical approaches to this step. The first is to analyze test scores descriptively for varying groups (e.g., interrupted versus not interrupted, crossing of interruption groupings with demographic groupings). This is the least rigorous of the approaches, but is likely to be the easiest to communicate to lay stakeholders.

The second approach is to conduct separate inferential analyses for every comparison of interest. This approach is typically carried out in combination with mechanisms for causal modeling of "experimental" effects in a context in which non-random assignment was used (e.g., propensity score matching). This is a more rigorous approach, but does not allow for controlling for one set of effects in modeling another. It also provides a middle ground in terms of the complexity of presenting and explaining results.

[7] Sinharay et al. (2015) do suggest one total score approach for investigating student-level impact (see p. 87).

[8] This can be treated as a continuous variable or can be broken into levels for the purposes of analysis.

The third approach is to create a statistical model in which all comparisons of interest can be conducted simultaneously and the structure of the model reflects the structure of the data.[9] This approach is also typically carried out in combination with mechanisms for causal modeling in a context of non-random assignment. This is the most rigorous approach least likely to be compromised by the weaknesses of the other approaches, but presenting and explaining the results is the most complex. If this approach is taken, great care will be needed to present and explain results in an understandable manner without sacrificing the accuracy gained by the complicated modeling approach.

Once effects on scale scores have been established, effects on derivative scores (e.g., performance level, growth, and growth level) can also be investigated. The simplest approach to investigating effects on derivative scores follows:

1. Create hypothetical scale scores for each student subject to one or more interruptions (by subtracting from the observed scale score the interruption effects identified in analyses of effects on scale scores).
2. Calculate hypothetical derivative scores from the hypothetical scale scores.
3. Investigate differences between observed derivative scores and hypothetical derivative scores.[10]

**Step 4A** is intended to identify effects of interruptions on item scores. There are three reasons for conducting such analyses. First, if there are not many items to analyze (as would be the case in fixed-form testing), performing item level analyses as a first step could be important in informing what effects to include in student-level analyses. The rationale for this is that overall student scores are somewhat muddied by the inclusion of both items from before interruption and after interruption. The use of a proxy variable (proportion of items reasonably assumed to be affected by a slowdown) may address this issue to some degree, but item-level scores are much more cleanly classified as having been affected by a slowdown or not. Item-level analyses are likely to be more sensitive and should assist in identifying effects to place in student-level models. If there is a large pool of items, this will likely be infeasible on a reasonable timeline because of the many thousands of statistical models needing to be estimated and interpreted.

Second, item-level analyses can be important in testing programs with relatively few responses per item and in which interruptions are widespread. In such cases, it may be desirable to retain item scores even from students whose overall scores are identified as potentially affected by interruptions. This may be necessary because of the need for a critical mass of item responses for use in post-administration analyses such as calibration, equating checks, and item analyses to support item review and/or item retirement. If this need is not critical, we recommend against attempting to parse out which items

[9] e.g., a multilevel model allowing for responses to interruptions to vary by demographic group, school, and/ or district.

[10] This is essentially calculating a counterfactual score for a given student. That counterfactual score represents **either** *what score would the student have achieved had no interruption been experienced*? **Or** *what score would the student have achieved had no interruption been experienced assuming that his or her response to interruptions was not idiosyncratic* depending on analytical design? This is a causal question, and the degree to which such an interpretation is supportable is a function of the care with which the analytical model attends to making causal claims. See also footnote 4.

tended to be affected and which did not tend to be affected, but to invalidate the item scores of all items administered to a student (whose score has been flagged as potentially affected) after that student experienced the interruption leading to his or her score being flagged.

Third, analyses of effects of interruptions on item scores can be useful for triangulating analysis of student-level scores and for investigating characteristics of items most likely to be affected by interruptions to inform any future analyses of similar events.

The second and third types of analyses may be simplified by limiting analyses to the effects identified in analyses of student scores.[11] Finally, the same three levels of rigor as described for analyzing student scores are applicable in analyzing effects on item scores.

**Step 4C** is intended to identify effects of interruptions on aggregate scores (e.g., mean scale scores) and their derivatives (e.g., percentages in performance level, mean growth scores, percentages in growth classifications).

As described above, through interruption effects on non-completion, aggregate scores may also be affected if non-completion is related to test performance. Identifying for which aggregate scores it is reasonable to suspect such effects is important.

There are two additional avenues for interruptions to affect aggregate entity scores, (1) through the effects on individual student scores, and (2) through entity-level effects not detected in the analyses of student-level scores.

The effect of the first avenue can be evaluated by creating hypothetical aggregate scores that would be expected in the absence of interruptions.[12] This can be done by replacing affected observed student scores with hypothetical student scores described in step 4B and re-aggregating results. Differences from observed aggregate scores to hypothetical aggregate scores can then be analyzed to flag aggregate scores likely to have been affected by slowdowns to a practically significant degree.[13]

We do not present recommendations on the second avenue through which interruptions could affect observed aggregate scores, as it seems unlikely that practically important effects not observed at the student level would appear only at the aggregate level.[14] The benefit of performing such analyses does not seem to be worth the cost in terms of the additional time required to complete the work when timely completion is at a premium.[15]

[11] However, there may be idiosyncratic effects on scores of specific items (i.e., effects that do not show up in analysis of overall student scores).

[12] See footnote 11.

[13] Statistical tests of significance can also be used, but with large sample sizes in state data sets and in some districts and/or schools, some statistically significant results are likely to be practically non-significant. We recommend flagging based on meeting the commonly accepted threshold for small effect sizes.

[14] Or that enhancement or muting of student-level effects at the aggregate level is likely to be practically important.

[15] However, this may be an interesting direction for further study after urgent work of identifying critical effects in a timely manner has been completed.

## JUDGMENT AND REPORTING

**Steps 5A-5C** and **Steps 6A-6C** involve judgment that can only be made by policy makers and other stakeholders, taking into account the evidence accumulated in all of the other steps and individual needs for reporting or otherwise using potentially affected item, student, and aggregate scores. We recommend that for the purposes of transparency, every item score submitted after an interruption, every student score contributed to by at least one affected item score, and aggregate score contributed to by at least one affected student should be flagged. In addition, we recommend that item, student, and aggregate scores should all receive an additional flag if they were subject to any type (or combination of types) of interruption identified as having an effect on those scores. Finally, we recommend that aggregate unit scores should be flagged for any degree of non-completion reasonably suspected to have an effect on aggregate scores.

While reporting adjusted scores is a risky enterprise, it may be necessary to do so in some cases for policy reasons. If state leaders decide there is a need to report adjusted scores, then care needs to be taken that such scores do not give a false impression of confidence. Therefore, in such cases it is important that not just the score be adjusted, but that corresponding increases in uncertainty be communicated.[16] This additional uncertainty introduced by score adjustment is important to communicate in scale scores, derivatives of scale scores (performance level, growth, growth category), and aggregate scores based on student scores.[17] There are two broad options for appropriately reporting adjusted scores:

1. Report the adjusted score and the upper and lower bounds of confidence intervals including both the original confidence interval and additional uncertainty attributable to score adjustment.

2. To minimize the potential of misuse, actual adjusted scores could be excluded from reports in favor of reporting only the upper and lower bounds of confidence intervals.

**Step 6A** is generally only applicable to determining what item scores not to carry forward into post-administration item-level psychometric activities. To maintain the psychometric integrity of the assessment, the easiest approach is to eliminate all item responses potentially affected by interruptions if it is feasible to do so. However, this may not always be the case for a small testing program in which field test or operational item response data is at a premium for program maintenance. In this case, the results of careful item-score analyses may be used to identify potentially affected item scores that are unlikely to have been affected by interruptions.

[16] Observed scores often have a standard error of measurement or a confidence interval reported with the scores. Adjusted scores may need to be reported with an expanded standard error of measurement or confidence interval reflecting both the uncertainty in those scores that arises through normal test administration and the additional uncertainty about the precision of score adjustment. For scores typically not reported with an associated measure of uncertainty, a measure of uncertainty may need to be introduced to reflect imprecision in score adjustment.

[17] For example, an additional "margin of error" around adjusted aggregated score reports based on the effects of interruptions on non-completion could be computed based on the number of additional students required to bring that school's or district's completion rate up to some acceptable threshold of participation/completion. The lower bound of this margin of error could be established based on average scale score (or percent in proficiency level) if all of those additional students scored at the midpoint of the not proficient category, and an upper bound established based on all of those students scoring at the midpoint of the proficient category.

However, if item scores are reported for individual students, the discussion for steps 6B-6C are also relevant.

Finally, **Steps 6B-6C** should be conducted regardless of the outcome of any other step. This step is intended to provide guidance for interpreting flagged scores. Each affected score may have one or multiple flags. For any flagged score that is reported, users should have immediate access to the cautions for interpreting flagged scores (e.g., using hover-over on a computer-delivered report, using footnotes on a paper report). For any adjusted score that is reported, additional cautions should be provided.

## Implications for Reporting and Accountability

In previous sections we defined technology-related interruptions, overviewed past studies that have examined these interruptions, and presented methodological approaches to examine the extent to which these interruptions may impact student performance. In this section, we examine options for dealing with the impact of interruptions. We will also explore the implications of these decisions on state reporting and accountability systems.

### OPTIONS FOR STUDENT SCORES

Once a state has determined that student scores are affected by the technology-related interruption(s), it must then figure out how those scores should be treated. There are several options for the treatment of student scores, ranging from doing nothing to suppressing the score. In deciding the best alternative, it is important that the guiding decision be made in terms of the option that does no harm to students, educators, schools, or districts. Below, we offer some alternatives and identify potential benefits and risks associated with these decisions.

*Do Nothing.* The state may decide to score and report student results in the usual manner. This may be a plausible option if the analyses reveal that there was little to no impact. Some states may also find this option attractive if the evidence reveals that the interruption improved student performance.

The obvious benefit of this approach is that it has the fewest immediate implications for the state's reporting and accountability system. However, there are a number of real risks to this approach. First, although overall results may suggest the impact is not systematic, that does not mean that some students or schools are not adversely affected if the results are used for accountability. Additionally, if the decision is made to move forward based on a perceived benefit to some students, it may be viewed as disadvantageous to other students. Finally, any adverse or beneficial impact that is ignored erodes longitudinal comparability for future administrations. In other words, it will be difficult to discern if the results the following year reflect changes in student performance when outcomes will be confounded by potential effects in the prior year.

*Estimate Student Scores on Unaffected Items.* If the interruptions affected only a few items, it may be possible to estimate the students' scores on all other items. The state may elect to estimate all student scores using the unaffected items. Alternatively, the state may choose to estimate the scores of only the affected students using the unaffected items.

The benefit of this approach is that students will receive scores which minimizes near-term impact on reporting and accountability. However, this alternative should only be used if the underlying test construct is unchanged by removing the affected items from the score. Another risk is that the resulting score will be somewhat less precise than the score based on all items. Therefore, it is important to evaluate the resulting precision along the full scale (i.e., conditional standard error) for assessments that are adjusted in this manner to ensure it comports with established thresholds. Involving the state's technical advisory committee in this process may be advisable.

*Adjust Affected Student Scores.* The state may be able to estimate how much student scores were affected by the interruption. In this case, the state could make an adjustment to the student scores to offset the estimated impact of the interruption.

Again, the benefit of this approach is that it provides a path forward to report scores. Moreover, it attends to the estimated impact of the disruption. However, is unlikely that all students were evenly affected by the interruption, so making a single adjustment may work well in some cases but not others. More complex or conditional adjustments may be difficult to implement and explain. Another risk is that the adjustments themselves will introduce some error in the outcomes, which may add uncertainty to accountability, uses, and longitudinal comparability.

*Give Credit for Affected Items.* The state may choose to give full credit for those items affected by the interruption. Credit for these items could be given to all students or to only those students affected by the interruption.

The primary benefit of this approach is that it preserves the ability to report scores and directly addresses the threat that the interruption may depress performance. Moreover, the approach targets items determined to be problematic. However, the benefit of guarding against a potential decrease in performance is also a risk, given that the adjustment is just as likely to artificially increase scores for some students. As with some other approaches described, this may serve to inflate error and erodes longitudinal comparability. Lastly, it should be noted that this approach is of limited utility when the impact of the interruption cannot be clearly isolated to a small number of items.

*Invalidate Scores.* The state may choose to invalidate the scores of all students who were affected by the interruptions. This typically involves a decision to suppress student level scores; however, there are a range of alternatives that are discussed more fully in the subsequent section.

Invalidating scores is often viewed as an attractive option when the state desires to act in abundance of caution to prevent a potential detrimental impact from interruptions. Additionally, it prevents misuse or misinterpretation of results that may have been adjusted. Of course, if scores are invalidated, then the state will need to consider implications for reporting and accountability, which is addressed in the subsequent section. The implications may extend beyond the current year. For example, most states now produce measures of academic growth which require one or more priors. Invalidating scores would impede growth calculations for at least two years.

## INVALIDATION ALTERNATIVES

Ultimately – and in some cases irrespective of the data – there are times where policy makers will decide that results should be invalidated. It is worth noting that there is more than one way to implement a decision to invalidate scores. Most commonly, this means that scores will not be produced at all, with student scores being treated as missing in district or state data files. However, other alternatives are available as reflected in table 3 below.

*Table 3. Alternatives for Handling Invalidated Records for Reporting and Accountability*

| | Report Student Level Result | Include in Aggregations | Include in Accountability Determinations | Mark Irregular/ Treat as Exceptionality |
|---|---|---|---|---|
| Complete Invalidation | No | No | No | No |
| Student Report Only | Yes | No | No | Yes or No |
| Student and Summary Report Only | Yes | Yes | No | Yes or No |
| Student, Summary, and Accountability Use with Caution and/ Adjustment | Yes | Yes | Yes | Yes |

In the first case, complete invalidation, a score is not produced and cannot be used for any accountability purposes. The advantage of this approach is that it eliminates a risk of stakeholders using the results in a way that may not be advisable. Certainly, this is an attractive option when scores are clearly impacted, particularly if the evidence suggests that the interruption had substantial influence on performance, whether that influence was unfavorable or favorable. However, some subgroups or educational entities may experience more interruptions than others, so care needs to be taken when invalidating scores (we address this issue in the Representativeness section below).

In cases where the impact is less certain, other alternatives may be appropriate. Such alternatives may be particularly appealing when one takes into account the understandable frustration felt by stakeholders if scores are not reported. That is, students, parents or teachers may feel they are being penalized for circumstances outside their control if scores are withheld. At the same time, it is understandable that education leaders will want to cast a broad net and invalidate any scores potentially impacted, even if the evidence is uncertain or weak. Such decisions may be made in abundance of caution, reflecting a desire to mitigate the likelihood of score misuse or errors in accountability outcomes. One way to 'thread the needle' may be to produce student level scores but omit the scores from aggregations, accountability outcomes, or both. Additionally, the score may be marked as 'irregular' indicating that it should be interpreted with caution. It may also be possible to use the scores in aggregate reporting or for accountability purposes but mark, and/or handle the instance as an exception. For example, aggregate scores could be reported as a range governed by a confidence interval to reflect uncertainty in outcomes. An example of a policy response is to invoke a 'hold-harmless' rule that stipulates the school's rating will not be downgraded in light of the ambiguity.

## ACCOUNTABILITY IMPACT

When a decision is made not to produce a score or to withhold scores from accountability uses, this raises the question of whether or how accountability decisions can be made. Fundamentally, this is a question about the impact of missing data. Missing data in accountability system (e.g. school or educator evaluation) is non-trivial as it can impact the precision and stability of either status (e.g., percent proficient, mean scale score) or growth analyses (e.g. VAM, SGP) and can introduce systematic bias in the resulting estimates (Braun et al., 2010).

There are two primary concerns with missing data in accountability, representativeness and precision. We explain each below and suggest analyses to investigate the extent to which each is potentially problematic.

### *Representativeness*

Representativeness refers to the extent to which a claim may be made that the performance of groups (e.g., subgroups, classes, schools) with missing or incomplete data differs from that which would have been observed if data were not missing. Two approaches for empirically addressing this claim are profile comparisons and predictions.

Profile comparisons refer to descriptive analyses to compare factors thought to be associated with achievement within year and across years. A within year comparison involves simply reporting side-by-side descriptive statistics for the cases that are included and excluded in the current year. Descriptive comparisons might include elements such as percent in ethnic group, economically disadvantaged, students with disabilities, and English language learners. While performance data for the missing tests will not be available, it is possible that measures do exist for an assessment that is well-correlated with the missing tests (e.g., a state norm-referenced test in a related content area). If this is the case, comparing included and excluded students on this measure adds another data point to compare the groups. It is important to do this for all levels in the system. For example, difference may not be detected at a district level that would be detected at a school level. Also, it is important to note whether certain subgroups are removed from the analyses due to n size rules, which can impact overall comparability.

Another approach for profile comparisons is to analyze outcomes from the previous year, removing cases for students excluded in the current year. That is, what results would have been produced in the preceding year(s) if the cases removed in the present year were missing? Not only can one review results to see if the remaining students 'look-like' the rest of the school on key elements, one can also observe the extent that performance (e.g., status and growth measures) were different with the missing cases in the previous year. If a retroactive look reveals the impact was substantial, that lends evidence to the claim that removing the cases would introduce a systematic bias in the current year.

### *Precision*

Precision refers to the extent that accountability scores are sufficiently accurate and stable with missing data. While a lack of stability alone does not necessarily indicate that the outcomes are not trustworthy, large shifts beg the question of whether the interpretation is due to fluctuations in student achievement or is explained by the introduction of error. For this reason, analyses to gauge the extent to which shifts in sample characteristics are thought to influence outcomes are important.

One approach is to examine multi-year performance. A straightforward approach is to produce descriptive statistics of group performance over multiple years at all levels of interest (e.g., school, class, subgroup). Comparing outcomes for the current year with previous years, as by visually inspecting outcomes on plot of multi-year results, provides some indication of stability. More formally, one can use analytic techniques such as regression to produce a predicted accountability score based on prior year scores. The outcome of interest is an observation of how many cases (e.g., schools) with missing data fall outside the prediction range compared to schools without missing data and the direction and magnitude of the shift.

Finally, there are a number of approaches to estimate the impact of sampling error that are useful to compare the precision of current year results to that of prior years. Hill and DePascale (2002) present four approaches to estimate the consistency of accountability classifications. One approach is termed split-half and simply involves dividing the data for each school into randomly equivalent halves and calculating the percentage of times the same decision is made for each half. Another method involves taking random draws with replacement by repeatedly producing random samples from the schools to evaluate decision consistency. The Monte Carlo approach can also be implemented, which involves simulating the distribution of scores and creating randomly generated samples from which classification consistency can be evaluated. Finally, direct computation involves calculating exact probabilities for correct classification by determining the distribution of errors. For an extended treatment on these methods, the reader is referred to Hill & DePascale, 2002.

In sum, these methods can be used to investigate the effect of missing data on the precision of the model by offering a means to compare decision consistency with full data in the prior year to outcomes with missing data in the current year. By so doing, one is able to evaluate whether and to what extent missing or incomplete data influence the precision of the model.

## Summary and Conclusion

The expansion of computer-based assessment has led to a predictable rise in testing disruptions due to technology failures. We acknowledge that a range of errors or disruptions can occur which lead to questions about the impact on student performance. Previous research in this area sheds light on a number of promising methods to investigate impact, which are detailed in the appendix of this paper. Building on prior research and applying new insights, we present a framework to guide investigations.

We outline some response alternatives in this paper and consider implications for accountability; however, policy makers may necessarily pursue alternatives that are not 'data driven.' For example, in most situations where a disruption occurs, policy makers will likely prioritize a response that minimizes a real or perceived negative impact to the students and schools affected. This is reasonable and explains why analysis that may reveal a minor or uncertain result may lead to a decision to invalidate results. Additionally, results that point to an improvement in performance of disrupted students would not likely result in a decision to raise performance expectations, whereas a similar finding in, for example, a mode comparability study could lead to production of different score tables. In the end, data analysis can play an important role in illuminating an appropriate response to testing disruptions, but it inevitably interacts with and is likely to subjugate to policy considerations.

## References

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington DC: AERA. Retrieved March 21, 2014, from http://www.apa.org/science/programs/testing/standards.aspx.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). *Getting value out of value-added. Report of a workshop*. Washington, DC: National Research Council, National Academies Press.

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 National Conference on Student Assessment, New Orleans, LA.

Hill, R.K., & DePascale, C.A. (2002). *Determining the reliability of school scores*. Portsmouth, NH: The National Center for the Improvement of Educational Assessment Inc.

Hill, R. (2010). *Reporting on Wyoming's testing irregularities.* Dover, NH: The National Center for the Improvement of Educational Assessment.

Hill, R. (2013a). *An analysis of the impact of interruptions on the 2013 administration of the Indiana Statewide Testing for Education Progress-Plus (ISTEP+).* Dover NH: The National Center for the Improvement of Educational Assessment.

Hill, R. (2013b). *Further analysis of gains made by interrupted students.* Dover, NH: The National Center for the Improvement of Educational Assessment.

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: Summary of Previous Studies

| Study | Methodology | Results |
|---|---|---|
| Hill (2010) | • Standardized mean differences of students experiencing interruptions from the state means between two years with latter year representing occurrence of administration problems<br>• Descriptive statistics, correlation coefficients, and significance testing<br>• By grade and all grades in reading and mathematics | • Standardized mean differences between two years were small<br>• Correlation coefficients of performance of students affected between years was substantial (0.72 in reading for 119 students and 0.78 in mathematics for 53 students)<br>• Mean differences of standardized scores were small (in reading ranged from -0.572 to 0.227; in mathematics ranged from -0.426 to 0.443)<br>• Mean differences in scaled scores by grade for reading, mathematics, and science were small or trivial<br>• None of the mean differences was statistically significant |
| Hill (2013a) | • Identified students affected two ways: (1) based on testing vendor information and (2) review of the list of students from vendor by local school systems.<br>• Counts by grade by type of school by length of interruption, number of interruptions, and test (math, ELA, science, and social studies)<br>• Comparison of mean scaled scores statewide by grade across five years for mathematics and English language arts<br>• Comparison of mean differences in scaled scores statewide between prior year of interruption and year involving interruption<br>• Comparison of mean differences in school means by grade by the percent of students interrupted in a school between prior year of interruption and year involving interruption<br>• Comparison of mean changes in school means by cohort by the percent of students interrupted in a school between results from two years prior and prior year of interruption and additionally between results from the prior year of interruption and year involving interruption<br>• Comparison of scaled score changes by student cohorts by year of testing and whether reported as interrupted for public and non-public schools<br>• Comparison of 2013 scaled score changes from previous year by student cohorts by test of first interruption (none, math, or ELA) for public and non-public schools<br>• Comparison of 2013 scaled score changes from previous year by type of interruption (none, any, timing of first interruption, and multiple interruptions within one session) | • Mean scaled scores statewide increased over a five year period for each grade except for a decrease of 1 point in grade 6 mathematics in the year of the interruption. Gains were not higher than historical patterns<br>• Mean scaled scores by grade from the prior year to the year of the interruptions did not change substantially except for grade 4 where, due to a policy change students not passing the grade 3 received grade 3 instruction and tested as a grade 3 student again instead of taking the grade 4 test<br>• Mean changes in student scores by the percentage of interruptions in a school (none, 1-19%, 20% plus) by grade did not show any discernable patterns in score changes. Mean differences between schools with interruptions and those without were about 1 point on a test where the student-level standard deviation is between 50 and 75 points<br>• Changes in scaled scores of cohorts of students (e.g., grade 3 to 4) in schools with no interruptions did not have larger gains than schools that were interrupted, and schools with more moderate amounts of interruption did not have larger gains than schools with larger percentages of interrupted students<br>• Cohorts of public school students had gains in the year of the interruption in three grades when the same students were compared to their prior grade performance and smaller gains in the two other grades for both mathematics and English language arts<br>• Students who were interrupted scored at about the same level, and often slightly higher, than students were not interrupted at all |

| Study | Methodology | Results |
|---|---|---|
| Hill (2013b) | • Comparisons of the mean scaled score gains of students in four groups representing (1) students in schools where no one at their grade was interrupted, (2) students in schools where other students at their grade were interrupted but not themselves, (3) students who were not reported as interrupted by the test vendor but were reported locally as having been interrupted, and (4) students who were reported as interrupted by the test vendor.<br>• Comparisons were made by the means and standard deviations of the mean scaled score gains for the same students across two years by content area, by comparison group indicated, and by socio-economic status represented by free or reduced price lunch.<br>• Comparisons on changing behavior of students' answers from wrong to right at various times when the change occurred with respects to the interruption (before or after) and when the item was actually presented to the student with respect to the interruption (before or after).<br>• The percent of items changed from wrong to right was calculated and compared by content area, by grade, and by when during the test the interruption occurred.<br>• Additionally, the impact on the raw score was calculated by students having made that change.<br>• Evaluations were made by comparing the percentages of items changed from wrong to right and the impact on the raw score. | • Mean differences between students with interruptions and students without interruptions were small (-3 to +5 scale points)<br>• Differences in standard deviations between students with interruptions and students without interruptions were small (0 to 3)<br>• Some students changed some of their answers when they returned to the test after the interruption and scored slightly higher than they would have scored without the interruption. However, this happened so infrequently that the impact overall was negligible. The raw score average increase ranged from 0.00 to 0.31. |

| Study | Methodology | Results |
|---|---|---|
| Bynum et al. (2014) | • Used student-level data to match students who experienced interruptions to students who did not by using propensity score matching involving demographic variables, previous year scores, and school-level achievement.<br>• Using these matched students, they compared the mean differences and distributions of differences between the two groups of students and then used regression to evaluate the effect of the interruption on the test scores by content area and by grade.<br>• Additionally, school-level aggregated data were used to evaluate the difference in mean test scores with and without the student scores who experienced the interruption.  Using regression, the authors modeled the effect of the test interruption on the school-level means and examine the impact on percent proficiency if predicted scores were used in place of the observed scores. | • Standardized mean differences between students who experienced interruptions and those who did not were small and not statistically significant, except for grade 4 in reading and mathematics, but the effect size was small.<br>• Disruption did not contribute to the predictability of the scores for the testing year when the interruptions occurred beyond prior year achievement, free/reduced-price lunch status, LEP status, race, ethnicity, gender, school-level achievement, and school-level percentage of free/reduced lunch.<br>• Scores from the year involving the interruptions were predicted using a number of variables for students who experienced a disruption and the same was done for a matched set of students who did not experience an interruption. The R2 difference indicated higher values for the students who did not experience an interruption. But, the difference in R2 was small (0.3% to 5%) by grade and test.<br>• Students' scores were impacted by the interruption in grade 6 reading (advantaged), in grade 8 reading (disadvantaged), in grade 3 math (advantaged), and in grades 4, 5, and 6 math (disadvantaged). These inconsistent results were small, affecting 10 to 40 students depending on grade.<br>• The impact of disruption on school-level scores was very small. Direction of the impact was inconsistent across grades and content area.<br>• Relationship of school means between prior year scores and scores in year when interruption occurred was not altered by the known prevalence of the disruption within schools. |

| Study | Methodology | Results |
|---|---|---|
| Sinharay et al. (2014) | • Student-level data were used to make comparisons of students' scores from interrupted and uninterrupted examinees by using (1) propensity score matching, (2) linear regression, and (3) IRT-based methods. | • Mean scaled scores and passing rates for five years were mostly as expected except for English and math grade 6 having lower 2013 mean score and math grade 3 having higher mean score.<br>• Comparison of students who experienced an interruption to a matched set of students based on propensity score matching, there were not statistically significant differences with small to trivial effect sizes by content area (English and math) and grade (4, 6, and 8). There was a statistically significant difference in grade 4 science, but the effect size was small.<br>• A linear regression predicting the scores from the year when the interruptions occurred using the prior year scores, gender, ethnicity, indicator of limited English proficiency, indicator of free/reduced-price lunch, and the interruption indicator showed no statistically significant difference based on the interruption indicator.<br>• The Bonferroni-adjusted 95% confidence intervals representing the mean difference between the actual scale score for each interrupted student and the matched uninterrupted student showed no difference.<br>• IRT-based comparison of the interrupted and matched uninterrupted students did not show a statistically significant difference for any content area and grade, except for grade 4 science where the effect size was small. |

| Study | Methodology | Results |
|---|---|---|
| Sinharay et al. (2015) | • Four methods to examine whether the interrupted students' performance after the interruption was worse than that expected from other information were used:<br>1. The first used the Wald test statistic to compare the ability estimates of students using items before the interruption with ability estimates of students using items after the interruption.<br>2. The second approach involved the same as the first except the empirical distribution of the calculated statistic was used to evaluate the differences in the ability estimates (i.e., empirical Wald).<br>3. The third approach was called a simple regression approach. This approach involved (1) the calculations of separate estimates of ability using IRT based on the items before and after interruption for both interrupted and uninterrupted students, (2) a regression was used for uninterrupted students to predict the ability estimates after the interruption using the ability estimates from before the interruption, and (3) using the regression coefficients from regression of the uninterrupted students, the ability estimates after the interruption were calculated for the interrupted students. Comparisons were made between the calculated and actual ability estimates for the interrupted students.<br>4. The fourth method involved a multiple-regression approach. This approach was also utilized following the same approach as the simple regression approach except that other variables were added in the regression predicting the ability estimates after the interruption for the uninterrupted students.<br>• Additionally, the type I error rate and power of these methods using a simulation study were evaluated. | • Based on the simulations, the Wald test had the largest type I error rate than the other approaches, and not used further.<br>• Based on the simulations, the multiple regression approach was the most powerful followed by the simple regression with the empirical Wald least powerful, but the lowest Type I error rate.<br>• The three approaches found very similar results with approximately 5% of the examinees identified as being impacted more favorably by the interruption and a similar percentage identified as being impacted adversely. These represent a negligible overall impact of the interruptions. |

CCSSO
Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072