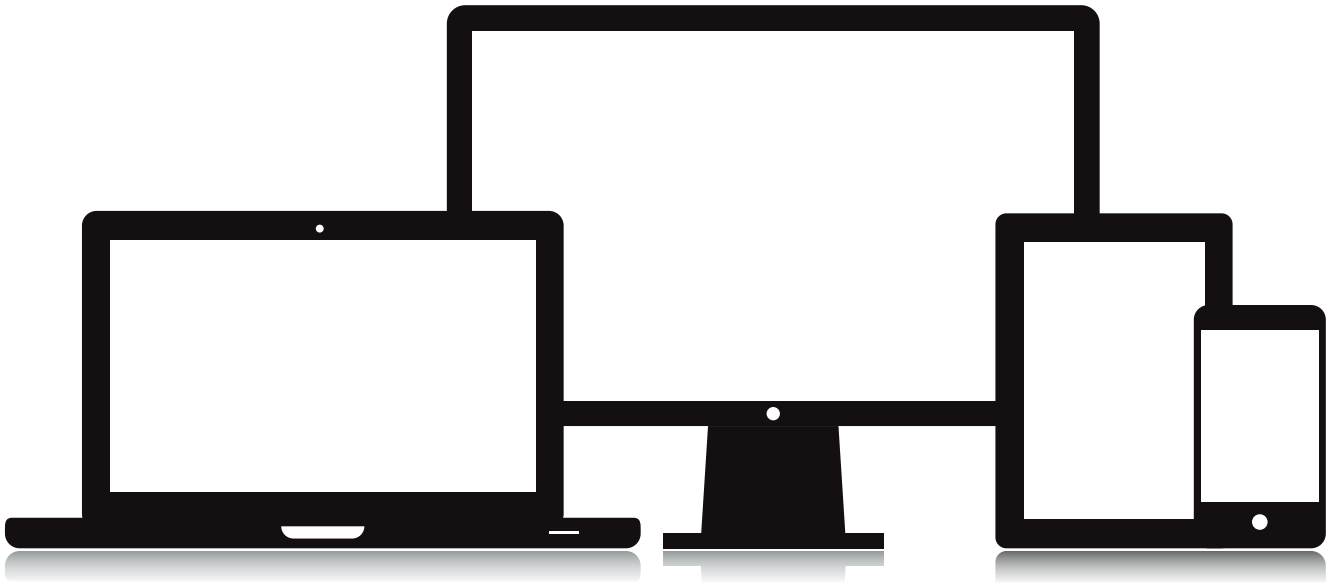
June 2016

Report commissioned by the Council of Chief State School Officers (CCSSO) Technical Issues in Large-Scale Assessment (TILSA) State Collaborative on Assessment and Student Standards (SCASS)



Score Comparability across Computerized Assessment Delivery Devices

Defining comparability, reviewing the literature, and providing recommendations for states when submitting to Title 1 Peer Review

Charlie DePascale, Ph.D.
Nathan Dadey, Ph.D.
Susan Lyons, Ph.D.

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.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS
Tony Evers, President, Wisconsin
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
Copyright © 2016 by the Council of Chief State School Officers, Washington, DC
All rights reserved.

Contents

Introduction

The shift from traditional paper-and-pencil testing to computer-based assessment may result in better information to *support learning and promote equity, and to measure progress and improve outcomes for our nation's students*, but it also introduces threats to the comparability of state assessment results across students, schools, districts, and states. As state assessments *move beyond bubble tests*, they are leaving behind many key contributors to standardization, including the No. 2 pencil and the bubble. For all of their faults and limitations, No. 2 pencils are ubiquitous and minimal training is necessary to correctly fill in a bubble on a paper test. In contrast, with large-scale, school-administered, computer-based testing it is inevitable that state assessments will be administered to students via a wide variety of technological devices. Variations in the manner in which test information is presented to students, and in the manner in which students interact with that information, must be carefully considered and accounted for in the design of assessments, in the production of student scores, and in the interpretation and use of assessment results. To ensure fairness, states must be able to confidently claim that the comparability of state assessment results is not impacted by variations introduced through the use of different types of technological devices to administer those state assessments.

In September 2015, the United States Department of Education (USED) issued non-regulatory guidance related to the Peer Review of State Assessment Systems (USED, 2015). Within this guidance, USED identified key areas in which states must be prepared to provide evidence to support claims of comparability across technological devices. With regard to test administration, the peer review requirements are rather straightforward. The state must provide evidence that they "…defined technology and other related requirements, included technology-based test administration in its standardized procedures for test administration, and established contingency plans to address possible technology challenges during test administration" (Critical Element 2.3 – Test Administration, p. 29). Examples of acceptable evidence include administration manuals or other key documents that "include specific instructions for administering technology-based assessments." These instructions should contain the actions necessary to ensure that test administrators and students are adequately familiar with the devices that will be used to administer the assessment. Although such standardized administration procedures may contribute to score comparability, they are not sufficient to ensure comparability. Therefore, the Peer Review Guidance also addresses the evidence that states must provide to demonstrate the comparability of results across technological devices. In short, the state must provide evidence that it has "followed a design and development process to support comparable interpretations of results" and "documented adequate evidence of the comparability of the meaning and interpretation of the assessment results" (Critical Element 4.6 – Multiple Versions of an Assessment, pp. 42-43).

For the specific case of technology-based state assessments, Critical Element 4.6 of the Peer Review Guidance requires the following:

> If the state administers technology-based assessments that are delivered by different types of devices (e.g., desktop computers, laptops, tablets), evidence includes:
>
> - Documentation that test administration hardware and software (e.g., screen resolution, interface, input devices) are standardized across unaccommodated administrations; or
> - Either:
>   - Reports of research (quantitative or qualitative) that show that variations resulting from different types of delivery devices do not alter the interpretations of results; or
>   - A comparability study, as described above.

The Peer Review Guidance further indicates that in the case where a state administers different versions of its state assessment (e.g., technology-based and paper-based assessments), the state must be prepared to provide results of a comparability study "that is technically sound and documents evidence of comparability generally consistent with expectations of current professional standards" (p. 43). Evidence of comparability is generally needed whenever there are variations in the content of an assessment or its administration – thus evidence is needed of comparability between pencil and paper assessments to technology-based assessment, as well as between the different devices a technology-based assessment is administered on (e.g., tablet to laptop).

The purpose of this document is to provide information and advice to support states in meeting USED Peer Review Requirements related to demonstrating the comparability of test scores across various devices used for technology-based testing. The document is divided into four main sections. In the first section, we discuss and define the comparability of scores. In the second and third sections, we present a brief summary of relevant research on the comparability of state assessment scores. The second section contains a summary of two key areas of score comparability research: 1) comparability between paper-and-pencil and computer-based assessments and 2) research on score comparability issues associated with student accommodations. The third section of the paper contains a summary of emerging research and analyses of comparability issues dealing directly with the use of various technological devices. The fourth section contains a synthesis and interpretation of the research findings issued related to score comparability across devices, as well as our recommendations for the type of documentation and evidence that states should be compiling in order to address those issues and present a solid case for peer review.

The document also includes an appendix that provides concrete steps states should be taking to minimize threats to comparability, document evidence of score comparability, and monitor potential threats to score comparability. The recommendations contained in the appendix are built on the information provided in the body of the document, but the appendix may also be used as a stand-alone document.

Section 1: Defining Comparability

The body of literature on the issue of score comparability is wide and varied, as are the definitions provided for score comparability itself.[1] One often cited definition of score comparability is that of interchangeability, that is "when comparability exists, scores from different testing conditions can be used interchangeably" (Bennet, 2003, p. 2; see also Winter, 2010; Way, Davis, Keng, & Strain-Seymour, 2016). Winter (2010) built on this definition and suggested that score comparability is specific to the *type* of score being used (e.g., an achievement level classification or scale score; see pp. 3-6). This means different scores created from the same test may have different degrees of comparability. For example, the classifications of students as above or below proficient may be comparable across two different versions of an assessment, while the full continuum of the scale scores may not be. For any given type of score, Winter (2010) notes that comparability requires that a "test and its variations must

- measure the same set of knowledge and skills at the same level of content-related complexity (i.e., constructs);
- produce scores at the desired level (i.e., type) of specificity that reflect the same degree of achievement on those constructs; and
- have similar technical properties (e.g., reliability, decision consistency, subscore relationships) in relation to the level of score reported" (p. 3).

Satisfying these conditions can be challenging, and there are numerous of sources of evidence that can potentially be considered to support score comparability. Recently, several authors have categorized approaches to defining score comparability and related methods of demonstrating comparability, including Sireci (2005), Abedi (2009), and Winter (2010). It is worth noting that these categories are inexact, as are all approaches used to classify the types of evidence needed to support score comparability. Sireci (2005) organizes approaches to establishing score comparability into five categories, which are based on predictive validity, structural relationships (e.g., factor analytic), test equating, networks of test-criteria relationships (i.e., nomological networks), and logical argument. Abedi (2009) defines six types of comparability: content and construct, depth of knowledge, accommodation, psychometrics, linguistics, and basic test features. In a similar vein, Winter (2010) suggests that comparability can be defined along two interrelated dimensions: content comparability (i.e., the assessed content) and score comparability (i.e., the type of score, such as an achievement level or scale score).

Our perspective is that each of these categories mentioned above summarizes types of evidence that can be used to support score comparability and that score comparability does indeed mean interchangeability. That is, comparable scores have the same meaning and can be used in the same way. Given the ways in which computerized devices have been used to administer the current wave

1 Due to the evolution of comparability research, at times the methods used to produce evidence of score comparability are equated with comparability itself (i.e., if the assessment data does not display differential item functioning across the test variations) then the scores are comparable. We take a different view in that score comparability can be supported by evidence of multiple methods and thus is not necessarily defined by a single analysis.

of state-mandated accountability assessments, the scores in question are almost invariably the scale scores. Given that scale scores are at a finer grain size than achievement-level classifications, showing the comparability of scale scores implies that aggregate scores or classifications derived from them, like achievement-levels, are also comparable.

To make claims about the comparability of scores across computerized devices, the comparability of the assessed content must first be established. This is most easily shown when the assessment items are identical across devices. Next, a planned series of research studies should be conducted to show that the testing administration device does not introduce construct irrelevant variance into the score estimates. Work on the comparability of scores produced under various accommodations provides guidance on how examinations of construct-irrelevant variance can be conducted and thus parallels can be drawn between score comparability across accommodations and score comparability across devices.

Section 2: Score Comparability Research related to Accommodations

Testing accommodations are "changes to a test or testing situation that are intended to improve student access to the content of the test without altering the test construct" (Pennock-Roman & Rivera, 2011, p. 10). Testing accommodations encapsulate a wide variety of methods designed to allow particular groups of students to demonstrate what they know and can do in ways that are not possible during typical administrations of the assessment. These groups of students are generally either students with disabilities (SWDs) or English Language Learners (ELLs).[2] Recent, key reviews of accommodations for SWDs include Laitusis, Buzick, Stone, Hansen, & Hakkinen (2012) and a long running series of reports supported by the National Center for Educational Outcomes (e.g., Rogers, Lazarus, & Thurlow, 2014; Rogers, Christian, & Thurlow, 2012; Cormier, Altman, Shyyan, & Thurlow, 2010). For ELLs, reviews include Pennock-Roman & Rivera (2011), Li & Suen (2012) and Kieffer, Lesaux, Rivera, & Francis (2009).

There are a wide number of accommodations, which should ideally be tailored to an individual student's needs (as argued by Pennock-Roman & Rivera, 2011, p. 11). Thus coming to a definitive conclusion about the comparability of accommodations for SWDs or ELLs in general is difficult. In addition, the requirement that the general and accommodated versions of the test measure the same construct, as noted in our quote by Winter (2010) above, can pose difficulties as accommodations often involve changes to the manner in which the construct is accessed or assessed. Such changes could result in a compromised measurement of the construct, however, without accommodations SWDs or ELLs may not be able to properly demonstrate the knowledge, skills, and abilities the construct represents. Given this tension, researchers have relied on what is known as the interaction hypothesis. This hypothesis states that an accommodation should improve test performance for the students who need the accommodation (e.g., SWDs or ELLs), but should not improve performance for those who do not need the accommodation (also see Way et al., 2016, for a similar high level overview of accommodations research).

2 That is not to say that SWD and ELLs are interchangeable, in fact appropriately assessing each group of students poses unique challenges and requires accommodations tailored specifically to that group.

Findings for some of the accommodations designed for SWDs have been consistent with the interaction hypothesis as shown in Laitusis et al.'s (2012) review, but the overall picture about many accommodations is decidedly mixed. For example, in the context of a read-aloud accommodation for English language arts, Laitusis et al. found that six of eleven studies investigating the interaction hypotheses did indeed support the hypothesis (see pp. 27-28). However, these studies differed in the way the audio content was delivered, what parts of the content was delivered via audio, and the student populations included in the study. These differences make it difficult to come to a general conclusion about read-aloud accommodations. Similar issues exist when trying to make a general conclusion about all accommodations for SWDs (see also Way et al., 2016). In contrast, the findings for many accommodations designed for ELLs have been consistent with the interaction hypothesis, as shown in Pennock-Roman & Rivera (2011) and Li & Suen (2012).

A general conclusion one might draw from research on accommodations is that although there is not a single universal conclusion – such as a conclusion that accommodations never impact comparability – there is a body of research that demonstrates cases and contexts in which the use of particular accommodations should not be considered a serious threat to comparability. Using those accommodations in unstudied contexts or introducing new accommodations, however, may pose a threat to score comparability. States and assessment developers can evaluate the context in which accommodations will be used on their assessment and draw on the relevant research to support the use of particular accommodations and determine those cases in which additional information is needed to support their use.

Section 3: Score Comparability across Devices

MODE COMPARABILITY

Research that has been done to investigate the comparability of test scores across paper-and-pencil and computer-based test administration can be considered a precursor to the current focus on device comparability. Unlike the focus on the interaction hypothesis discussed in research on accommodations, most studies investigating score comparability across modes of administration have focused on the interchangeability of scores between pencil-and-paper and computer-based test versions (i.e., would a student who took the test on a computer have received the same score had he or she taken the test using the paper version). These studies are generally aimed at investigating whether the difference in modes (paper vs. computer) causes different levels of performance. Thus investigations of the comparability of scores from paper-based and computer-based tests have been less focused on producing evidence of score comparability regarding the construct and technical properties, as outlined by Winter (2010), and more focused on producing evidence based similarities or differences in student performance. If students, matched on relevant characteristics, have similar levels of performance on both versions, the conclusion is that the scores are comparable and can be used interchangeably. Ideally these matched groups of students would be the result of random assignment of students to a computer-based

or paper-based test, and thus differences in performance can be directly attributed to the differences in mode (i.e., a mode effect). In operational settings, of course, such an ideal is not typically practical and other forms of matching students are applied.

Several recent meta-analyses (Wang, Jiao, Young, Brooks, & Olson, 2007; 2008; Kingston, 2009) have concluded that in general, mode effects between paper and computerized test administrations appear to be small (e.g., non-significant or small effect sizes). However, these meta-analyses do include a number of studies that do show mode effects. For example, of the 81 effect sizes examined by Kingston (2009, Table 1) more than one-third were fairly large (i.e., greater than 0.1), with 13% indicating that the computer test version is associated with lower performance relative to the paper test version, and 22% indicating higher test performance on the computer version. Kingston (2009) shows that these types of differences were partially explained by content area, in that the mode effects tended to favor computer administration in English langue arts and social studies and paper administration in mathematics.

Other meta-analyses that predate those above have come to similar conclusions (e.g., Mead & Drasgow, 1993). All of these analyses suggest that while mode effects are often small, there are numerous cases in which the effects are large, and that these cases are difficult to predict. Ultimately, the conclusions about mode effects in the literature are mixed (cf. Way et al., 2016). Throughout both current and more dated studies, numerous authors (e.g., Wang et al., 2008) note that the general findings of meta-analyses do not mean that testing programs can forgo conducting studies of comparability. In a review of comparability studies, the Texas Educational Agency notes that

> Because the majority of comparability studies in other states have found the computer- and paper-based versions of their tests to be comparable overall (see Tables 3–6), a natural question to ask is: have we amassed enough evidence (or will we ever get to such a point in the near future) to say that computer- and paper-based tests are comparable so that no more comparability studies are necessary? The answer depends on the specific needs and circumstances of each testing program. Each state needs to assess its situation and weigh the costs of conducting regular comparability studies against the risks of not conducting them (2008, p. 34).

In a similar vein, we believe that research on score comparability should change to better allow states to assess those costs and risks. Consistent with the manner that research has addressed the purpose of accommodations, comparability studies must more fully capture the current context. That is, the need to administer tests on different devices while fulfilling the purpose of testing: to arrive at the truest estimate of the examinee's score on the tested construct. To support this goal, studies should show that device effects are not introducing construct irrelevant variance. The next part of this paper briefly summarizes the current state of research evaluating the effect of computerized administration devices on score comparability.

SCORE COMPARABILITY ACROSS COMPUTERIZED DEVICES

Device comparability is a relatively new area of research due to the current increase in the use of laptops, tablets, and other mobile computing devices for the administration of assessments. Due to the relative nascence of this area of research, studies examining device effects are limited, and therefore the majority

of the work cited in this review results from an effort of gathering "gray literature"[3] from industry-specific analysis and practices. For the purposes of this paper, the authors contacted 17 testing organizations to gather studies in the areas of device comparability. Of those contacted, eight responded with reports of recent studies completed that specifically examine the comparability of scores resulting from different computing devices. These studies, along with the relevant literature cited by those studies, are included in this review. We have organized the studies into two broad categories. First, we discuss the results of studies that examine the effects of devices on score comparability (e.g., laptop vs. desktop, tablet vs. computer). Secondly, we summarize study results related to features of devices that have been shown to impact score comparability and should be considered when examining device effects (e.g., screen size, on-screen vs. external keyboard, and familiarity with device).

RESEARCH RELATED TO DEVICE EFFECTS

Some of the earliest device comparability studies examined differences in scores and student experiences between laptop and desktop computers. In 1996, Powers and Potenza at Educational Testing Service found no differences in GRE Verbal and Quantitative measures by testing device, and small differences in student writing performance on the essay prompts. The differences found were attributed to the possible difficulty of switching from a full-size desktop computer keyboard to the smaller keyboard on the laptop. At the time of this study, participants reported little prior experience with laptop computers. However, due to the rapid proliferation of laptop computers in the past 20 years, these findings may not generalize to today. More recently, a 2005 study of the National Assessment of Educational Progress (NAEP) also found that student writing performance was slightly lower for students using laptop computers than students writing on desktop computers (Sandene, Horkey, Bennett, Allen, Braswell, Kaplan, and Oranje, 2005). However, this finding was not consistent across essay prompts or studies, including a larger study finding no differences for males, but with female students performing significantly lower on NAEP-provided laptops than when using school desktops (Sandene et al., 2005). In 2015, Davis, Kong, and McBride from Pearson find that students may have a preference for taking tests on devices with which they are most familiar, but in their study of 934 high schoolers, this preference did not seem to translate into performance. This study examined the comparability of scores for students testing on computers to those testing on tablets and found no significant device effects, results which held across content areas and item types. A second 2015 Pearson study by Davis, Orr, Kong, and Lin confirmed these results by using an experimental design to test for score comparability across laptops and tablets (with and without external keyboards). Again, after controlling for prior achievement, no statistically significant effects were detected at any grade level. Though score differences were not formally examined, Yu, Lorié, and Sewall (2014) at Questar Assessment conducted cognitive laboratories where it was found that students generally experience more frustration responding to items on a tablet interface than on laptops or desktops. In this study, a higher number of students indicated that they would rather take a high-stakes exam on a laptop or desktop than the number of students who reported a preference for tablets. Lastly, one study was found that compared scores from a small,

3 Gray literature is a term used to refer to information found outside of traditional academic and published journals. Among other types of information, it includes internal reports, technical papers, and project reports prepared by organizations.

handheld Personal Digital Assistant (PDA) device (3.7 inch display) with assessment results from laptops. Schroeders and Wilhelm (2010) used confirmatory factor analysis to understand the factor structure of reasoning ability as measured by two devices. These researchers found small and uncorrelated device-specific factors for the PDA and the laptop. They hypothesized the factors may be attributed to differences in familiarity with device, item presentation modes, or differing motor skill or perceptual demands and suggest further research will need to investigate the plausibility of these factors.

FACTORS THAT MAY CONTRIBUTE TO THE PRESENCE OF DEVICE EFFECTS

Familiarity

Though we found little evidence that device comfort and familiarity has been formally studied, it is frequently cited as a potential threat to comparability (see Powers & Potenza, 1996; Sandene et al., 2005; Lorié, 2015; Schroeders & Wilhelm, 2010; Keng, Davis, McBride, Glaze, & Steedle, 2015). In his 2015 presentation to the Association of Test Publishers, Lorié stresses the importance of understanding the impact of devices on score comparability as context dependent. He argues that the degree to which scores are comparable is inextricably related to a set of skills he terms "device fluencies." Since device fluency is a prerequisite for appropriately accessing any computerized assessment, Lorié suggests that test takers should be tested on their device fluencies as part of the assessment domain to ensure they have the minimum required level to access the tested content of interest. He calls this "comparability by design." Davis and Strain-Seymour (2013a) investigated differences in device comfort and familiarity and found that exposure to devices varies by age and preference may vary by content. For example, eleventh graders reported using tablets in school, while fifth graders reported more frequent use of laptops over tablets. Additionally, the majority of students reported that they would prefer to write essays using desktops or paper over tablets. These preferences are likely related to what could be called device fluencies. For example, students have reported difficulty with scrolling on iPads as the scrollbar only appears once students have begun to scroll which requires a bit of prior knowledge or experimentation (Pisacreta, 2013). The next section of this paper discusses those additional device features that may comprise or require device fluencies.

Device Features

Screen Size

There are two separate but related issues to consider when evaluating the effect of the screen size of the test delivery device: 1) the physical size of the display, and 2) the amount of content shown at once on the display. From the research reviewed for this paper, it seems the latter issue is of greater concern. Results suggest that, holding the information shown on the screen constant, screens of 10 inches or larger are suitable for viewing and interacting with assessments, with little evidence of test performance differences or item-level differences (Keng, Kong, & Bleil, 2011; Davis, Strain-Seymour, & Gay, 2013). Evidence suggests that smaller screens may introduce challenges that threaten comparability (Davis, Strain-Seymour, & Gay, 2013; Schroeders & Wilhelm, 2010). However, it seems to be that the amount of content displayed on a screen

without the need to scroll can affect test performance, particularly for assessments requiring the reading of passages. Bridgeman, Lennon, and Jackenthan (2003) found that while mathematics performance remained stable across differences in computer monitor sizes, the same was not true for the assessment of verbal skills. When the percentage of the required reading material visible at any one time was reduced, verbal scores were depressed by about a quarter of a standard deviation. This may be because factual recall of textual information has been shown to suffer as the amount of scrolling necessary to read a complete passage increases (Sanchez & Goolsbee, 2010). Additionally, Davis and Strain-Seymour (2013b) found that features of the assessment (e.g., calculator tool or on-screen keyboard) that block part of the test content can add additional strain to the working memory. Sanchez and Branaghan (2011) confirm earlier findings that small screen sizes, necessitating additional scrolling, can interfere with recall, and even more seriously, reduce ability in complex reasoning. However, these researchers find that repositioning the small device to the landscape orientation effectively mitigated the negative effect of the small screen size, and this change in device orientation seemed to be especially beneficial for lower ability participants (Sanchez & Branaghan, 2011).

Input Mechanism

Touchscreen inputs are commonplace for tablets and smaller mobile devices used for testing, but do not come without introducing possible threats to comparability. In general, some input precision is compromised when using a fingertip rather than a mouse (Way et al., 2016). The most common issue with fingertip input is when objects in the screen requiring interaction (e.g., selection, drag-and-drop) are close in size or smaller than the student's fingertip, or when objects are close together (Strain-Seymour, Craft, Davis, & Elbom, 2013; Eberhart, 2015). Additionally, a mouse allows for "hovering," visually showing the cursor's location on the screen, which supports accuracy of input and guided reading (Way et al., 2016; Eberhart, 2015).

Keyboard

Studies that examine effects associated with on-screen keyboards, as opposed to external, find that they work equally as well for short or single-response items, but student responses tend to be reduced in length when using onscreen keyboards for responding to open-ended or composition items, likely due to fatigue (Davis & Strain-Seymour, 2013b; Pisacreta, 2013). Because students cannot rest their fingers on the onscreen keyboard, students' keyboarding skills are restricted and they instead defer to the "hunt-and-peck" method to input their responses to essay items. This typing method generally results in less accuracy and takes longer than traditional keyboarding, and Pisacreta (2013) found that most participants have a preference for an external keyboard when responding to essay prompts. However, the same preference was not found with younger students, who are less experienced typists (Davis & Strain-Seymour, 2013b). Additional considerations related to the use of an on-screen keyboard include device fluency features, for example, knowing how to switch between letters and numbers on the keyboard and facility with highlighting and moving texts, and screen real estate, in that the keyboard often uses valuable screen space, blocking test content (Pisacreta, 2013; Davis, Strain-Seymour, & Gay, 2013; Strain-Seymour et al., 2013). As previously discussed, device fluency and content display have both been cited as potentially interfering with score comparability. Again, it may be that device positioning could help to offset any negative effects associated with the on-screen keyboard. Yu et al. (2014) found that students who rated themselves as advanced users of the onscreen keyboard tended to position the tablet flat on a surface

while typing. Students who had beginning typing skills with the onscreen keyboard however, preferred to prop the tablet up at an angle. More research should be conducted to investigate whether laying the tablet flat may help improve the assessment experience when using onscreen keyboards.

Assessment-specific Features

Content Area

Though researchers at Pearson found consistent results across assessed content areas, both the STAR assessments from Renaissance Learning and the PARCC assessments find interactions between device effects and content. Renaissance Learning investigated score differences resulting from test administration through their iPad application as compared to test delivery via computer through their web application. STAR Early Literacy, which measures foundational skills related to reading, language, and numbers and operations, was found to show no statistically significant differences in performance across the two platforms. However, both the STAR Reading and STAR Math tests had results that, in some grade spans, favored the web application in small, but statistically significant ways (Renaissance Learning, 2013). It may be that the differences found across these three assessments have more to do with age of examinee than content area. As part of a robust field testing research agenda investigating comparability in multiple ways, device effects across tablets and computers for the PARCC assessment were detected and differed in strength and significance by both content area and grade level (Keng et al., 2015). As part of their field testing research, PARCC. studied device comparability for a sample of six assessments: grade 4 mathematics (Math) and English language arts/literacy (ELA), grade 8 Math and ELA, and grade 10 Geometry and ELA. Comparability was examined through item level analyses (comparing means and difficulty estimates), component-level analyses, and test-level analyses (comparing reliability, validity, and score interpretations). The item- and component-level analyses revealed evidence of score comparability across devices except in the cases of grade 4 Math and grade 8 Math, respectively. The evidence indicated that approximately 37% of the math tasks were more difficult on the tablet condition than on the computer for grade 4 mathematics. For the component analysis, the correlation between the PARCC end-of-year assessment the PARCC performance-based assessment is statistically lower for the tablet condition than for the computer condition in grade 8 mathematics. Test-level reliability analyses also revealed differences across conditions for the grade 8 mathematics exam (favoring the computer condition), and additionally, the grade 10 ELA assessment (favoring the tablet condition). Lastly, the convergent validity evidence and examination of comparability of raw score interpretations provided evidence of device comparability, except in the case of the grade 4 ELA assessment where the correlation with a criterion measure was weaker for the tablet condition, and the raw score concordance analysis provided evidence that the tablet condition was more difficult (Keng et al., 2015).

Item Type

Research indicates that the degree of score comparability across computing devices may vary with the types of tasks with which students are interacting (Eberhart, 2015; Davis & Strain-Seymour, 2013a; Davis, Strain-Seymour, & Gay, 2013). Eberhart (2015) examined differences in student performance across computer and tablet conditions for both math and ELA. In both content areas, test performance

slightly but significantly favored the computer condition; however, there were significant interaction effects between item-type and device. Though performance was higher for multiple choice items on the computer, the same effect was not detected for technology-enhanced items (Eberhart, 2015). Davis et al. (2013) conducted a series of "think-alouds" with students taking assessments on tablets that provide insight to help explain this interaction effect. Though, as mentioned above, students had some difficulty interacting with items when objects were small or close together, for items that were designed well for the tablet, students commented on the favorability of being able to directly interact with the item by using their fingertips to drag-and-drop objects. Bar graph questions seemed to work particularly well, as compared to Cartesian graphing questions that required the input of precise points (Davis, Strain-Seymour, & Gay, 2013). However, Davis and Strain-Seymour (2013) find that students who are familiar enough with tablets were able to overcome difficulties with precision by using the "pinch and zoom" method of enlarging the test content.

Section 4: Synthesis and Recommendations

In the first three sections of this document we discussed the meaning of score comparability, reviewed relevant historical research on the general issue of score comparability, and summarized emerging research directly addressing score comparability across computerized devices. In this section, we attempt to pull that information together to support states in their effort to define, discuss, examine, and draw conclusions regarding score comparability across devices within their state assessment programs. We begin by discussing a series of steps states can follow to clearly define issues related to the impact of the use of different devices on score comparability and also to isolate those issues from other threats to comparability:

1. Identify the Comparability Concern(s) Being Addressed
2. Determine the Desired Level of Comparability
3. Clearly Convey the Comparability Claim or Question
4. Focus on the Device

We conclude with a recommended approach that states can follow to gather and present evidence of score comparability across devices as part of their ongoing efforts at improving the quality of their assessments as well as to meeting the requirements of Peer Review.

Identify the Comparability Concern(s) Being Addressed

The focus of this document is on score comparability across devices, but it is likely that states will be addressing several comparability concerns within their state assessment program simultaneously. Critical Element 4.6 of the Peer Review Guidelines, Multiple Versions of a Test, contains two distinct comparability questions. Comparability across devices, of course, is one of those concerns. As implied in the title of the element, the second comparability concern is related to state assessment programs

that administer multiple versions of its operational assessment. *Multiple versions of a test* is a phrase that encompasses a wide variety of test designs, many of which might apply within a single state assessment:

- *Different operational test forms are administered to students within the same year*; that is, the items on which individual student scores are based vary across students. This condition applies to computer-adaptive tests (CAT) such as Smarter Balanced and also to fixed-form programs such as PARCC that administer multiple test forms within a single administration window.
    - Note that virtually all state assessment programs administer different operational test forms across years. Although this may not be regarded commonly as a comparability concern, the issues (as well as the solutions and required evidence) are similar.
- *A single operational test form contains embedded non-operational items* that vary across students. This condition applies to state assessment programs that contain embedded field test items or perhaps equating items that are matrix-sampled across multiple test forms. Although all student scores are based on the same set of items, students experience multiple different versions of the test.
- *A single operational test form is translated into different languages*. In this condition, all students encounter the same set of items, but those items (and related test materials) are presented in different languages.
- *A single operational test form is administered with a set of allowable accommodations.* This is the norm in most state assessment programs. At a minimum, students with disabilities participate in the state assessment using accommodations defined in their Individualized Education Program (IEP) and considered allowable by the state.[4] Some state assessment programs also offer accommodations within the standard form that are specifically designed for ELLs. Some state assessment programs make accommodations available or allow some flexibility in administration conditions (e.g., untimed or loosely timed tests) for all students.
- *A single test form is administered in different modes (e.g., paper-and-pencil and computer).* In this condition, the same test content is administered to all students, but students interact with the assessment (i.e., receive and respond to the test information) differently.

Therefore, the condition in which the same test content is administered to all students, but through the use of different technology-based devices (e.g., desktop computers, laptops, tablets) is only one of many ways in which a state may be administering multiple versions of its test.

At this time, it is commonplace for a state assessment program to include all, or most, of the conditions described above. In fact, most of those conditions reflect best practices implemented in order for the state to meet other Critical Elements for Peer Review, such as those related to Test Security (2.5) or Inclusion of All Students (5.1 – 5.4). Each of those conditions, including the use of different technology-based devices, presents the state with comparability concerns that the state must address individually and collectively. The steps taken to mitigate each of the threats posed to comparability and the evidence needed to demonstrate comparability vary across the conditions. It is critical for the state

4 This refers to students who will receive a score on the assessment that is considered comparable to scores of other students. There may also be students who participate in the state assessment using accommodations defined in their IEP who do receive a score that is not considered comparable because it modifies the construct being assessed.

to a) identify individual threats to comparability, b) develop a plan to mitigate each threat individually, c) identify evidence to document or support the steps identified in that plan, and d) identify and collect the evidence needed to demonstrate that score comparability exists collectively across all of the conditions that apply to the particular state assessment program. It is equally critical for the state consider the interactions among all of the identified threats to comparability and to consider how those threats fit within the larger context of the assessment program. Some threats to score comparability are introduced because of the need to offset other threats (e.g., security, fairness). Much like it is inappropriate to attempt to maximize reliability at the expense of validity, a state cannot set out to minimize or eliminate individual threats to score comparability without considering the collective impact of those actions.

Determine the Desired Level of Comparability

It is clear that states are interested in interchangeability of individual student scores on state assessments – the strongest level of score comparability. That is, states intend to report student scores on the same score scale and aggregate scores across tests administered on different devices with no regard to the device used to administer the assessment.[5]

Clearly Convey the Comparability Claim or Question

Simply determining what level of comparability is desired is not enough; additional information is needed to clearly convey the claim that the state is making about student performance on the state assessment. Different claims will lead to either different evidence being collected to support the comparability claim or different analysis and interpretation of the same evidence. Consider the following statements:

- If a student took the state assessment on another device, he or she would have received the same score.
- The student took the state assessment on the device most likely to produce the most accurate estimate of her or his true achievement.

The first statement makes a claim that student scores are device-neutral. In contrast, the second statement explicitly allows for the possibility that a student will perform better with a test administered on one device than another. Both claims, however, are acceptable within the context of score comparability. And both claims are commonly made within the context of standardized, large-scale assessments.

The first statement places a premium on the traditional concept of standardization. Everyone takes the assessment under the same conditions and intended users of the test results know the conditions under which the test was administered. States administering a college admissions examination as their state

5 This might not be true in the case where a state administers an alternate version of its state assessment for students who are not able to participate in the general assessment even with standard allowable accommodations. In that case, the state might be most interested in claiming comparability of performance level interpretations rather than interchangeable scores. These are students being measured against the same academic achievement standards as those students participating in the general assessment and not students with significant cognitive disabilities participating in the Alternate Assessment with Alternate Academic Achievement Standards (AA-AAAS).

assessment may be engaged in discussions about the administration conditions that support this claim. The premise is that all students must complete the assessment under a tightly specified set of conditions or their results cannot be used for a particular purpose, such as college admission.

The second statement contains an increased acceptance of flexibility and reflects the current practices within most assessment programs with regard to the use of individual student accommodations, which are offered to allow for improved access to the intended construct. As long as that flexibility in administration is determined to not alter the construct being measured and is consistent with the intended interpretation and use of the assessment results, differences in administration conditions are allowable and are, in fact, encouraged.

With regard to score comparability across devices, states are most likely to make a comparability claim that allows for flexibility, rather than one that argues that scores are device-neutral. That is, states will not claim that a student would have received the same score if he or she took the assessment on a different device. States are more likely to claim that allowing districts and schools to administer the state assessment on devices with which the students are familiar removes barriers to performance, thereby providing students a better opportunity to demonstrate what they know and are able to do. In summary, devices will not be regarded as interchangeable as if they were No. 2 pencils. Rather, states will and should accept that familiarity and fluency with a particular device used to administer the assessment is a factor that impacts the ability to produce the most accurate estimate of a student's true achievement.

This increased emphasis in flexibility is *not* coupled with a decreased need for comparability and evidence to support it. Either example claim, as well as any other claim a state might make, requires similar-levels of evidence. The difference is in what comparability evidence is collected as well as how the evidenced is used to support score comparability across devices.

Focus on the Device

When each of the issues described above has been addressed, the state will be in a position to focus its attention on issues directly related to the way in which differences among devices might impact score comparability. To be clear, questions about score comparability across devices are distinct from other threats to score comparability such as the following:

- differences in inclusion policies,
- differences in test administration procedures,
- differences in test content,
- differences in the types of items or the format of items used on the assessment, or
- differences in scoring and/or the response that a student is expected to provide.

Each of those factors could impact score comparability, but their impact is not limited to the use of different technology-based devices across students. In other words, with respect to concerns about score comparability, each of the above issues is likely to supersede issues associated with the use of different

devices. Administering multiple versions of a state assessment within a year that include any of the differences noted above or making changes to a state assessment program across years that include any of the differences noted above is likely to impact score comparability even if the same devices are used by all students within and across years.

Questions about score comparability across devices are likely to include concerns about differences among students in the following areas:

- the manner in which content is presented,
- the manner in which students interact with the content presented, and
- the manner in which students respond to the content presented.

As demonstrated in the research literature reviewed in this paper, there are key aspects in each of those areas above which have been identified as threats to score comparability.

When different devices are used, the state, at a minimum, should attempt to eliminate or minimize differences in each of the areas listed above. For example, the current literature suggests that differences in devices can be minimized if all students are sufficiently fluent with the functionality of the device on which they are testing; the amount of content that appears on the screen without requiring scrolling is the same across devices; the items are designed for comfortable use with fingertip input when touchscreen devices are used (e.g., items are large enough and spaced widely enough); and external keyboards are available for response to essay prompt. In short, states should identify critical aspects related to devices that are likely to impact score comparability and standardize those features across devices. When differences in those factors cannot be eliminated, the state should be prepared to present evidence that remaining differences do not negatively impact score comparability.

CONCLUDING RECOMMENDATIONS

Ultimately, the documentation and evidence that states should produce to meet Peer Review requirements for score comparability across devices should be consistent with the comparability claim that the state is making. In most cases, states will likely want to argue that the students are taking the state assessment under the conditions most likely to produce the most accurate estimates of their true achievement. Therefore, the type of evidence submitted should parallel that which is used to support the use of accommodations (Critical Element 5.3):

- Description of the reasonable and appropriate basis for the set of accommodations offered on the assessments, such as a literature review, empirical research, recommendations by advocacy and professional organizations, and/or consultations with the State's TAC, as documented in a section on test design and development in the technical report for the assessments.
- For accommodations not commonly used in large-scale State assessments, not commonly used in the manner adopted for the State's assessment system, or newly developed accommodations, reports of studies, data analyses, or other evidence that indicate that scores based on accommodated and non-accommodated administrations can be meaningfully compared.

- Evidence that the State has a process to review and approve requests for assessment accommodations beyond those routinely allowed, such as documentation of the State's process as communicated to district and school test coordinators and test administrators.

Commonly used devices

As with commonly used accommodations, for technology-based devices commonly used to administer assessments, states (and their assessment contractors)[6] should plan to draw on existing, available research on the comparability of scores when tests are administered using those particular devices. The state should be able to produce documentation to show that the use of the device does not negatively impact the comparability of scores. The state should also be able to produce documentation to demonstrate that it employed test design and administration specifications to ensure that the device was used in a manner consistent with previous use and best practices.

New devices

For newly developed or rarely used devices, an increased burden of proof to provide evidence of score comparability does fall on the state. If new devices are merely an adaptation of, or a variation on, existing devices, evidence should include documentation on similarities and differences with existing devices on key aspects and design features that have been shown to impact score comparability. In many cases, the introduction of new devices will represent merely incremental changes over previously researched devices and require minimal new evidence to support their use. In other cases, however, there will be significant changes such as the drastic reduction of screen size, the removal of external keyboards, and/or the introduction of touch screen capability. In accordance with best practices, states should not introduce such new devices into their state assessment programs on a wide-scale basis without a prior understanding of their impact on student performance and score comparability.

Program monitoring

For both commonly used and newly developed devices, the state should collect information on the devices used and develop a plan to monitor performance across devices on a regular basis. In accordance with best practices, even in cases where one expects there to be no threat to score comparability (e.g., field test items are embedded in test forms using a matrix-sampled approach, items are administered to students in a random or non-uniform manner, or accommodations are allowed), the state should conduct secondary analyses periodically to show that there is no negative impact on score comparability. As with the use of accommodations, such a monitoring program is particularly important when there is local control over the manner in which policies are implemented. With accommodations, states regularly monitor local policies and actual accommodations use to ensure that allowable accommodations are not being overused or underused by some school districts. Similarly, with the use of technology-based devices, the state should be aware of any systematic differences in the use of devices across districts and be prepared to investigate the potential impact of any such systematic differences that might be found.

6 Although the state bears ultimate responsibility for all aspects of its state assessment program, it is assumed that states' assessment contractors will be active partners in providing and generating evidence of score comparability related to the use of its off-the-shelf or custom-developed assessments.

A comprehensive view of score comparability

There can be little argument that the field of large-scale assessment, in general, and state assessments, in particular, is in a state of transition. Much of that transition, and the resultant changes to assessment programs, is intentional and directly related to advances in content standards, assessment policies, and available technology. The administration of state assessments on a variety of devices across schools and districts is just one of the consequences of that transition that might impact the comparability of scores across students within and across years. As stated above, a state should understand the potential impact of the use of different devices, attempt to mitigate that impact, and regularly monitor that impact. Those activities, however, should occur within a comprehensive assessment plan that comprises all aspects of the assessment program, including consideration not only of the interactions among all factors that might impact score comparability, but also includes consideration of how score comparability should be regarded by the state during and following the transition period.

There are few straightforward questions with black/white answers in large-scale state assessment. Decisions made to increase reliability might negatively impact validity. Decisions made to relax standardization and increase flexibility to enhance inclusiveness might change the claims that can be supported or interpretations made based on the assessment results. Similarly, decisions about score comparability will also come with tradeoffs. For the appropriate internal use of state assessments and to meet the requirements of Peer Review, states should understand and be able to provide a rationale for each of their design decisions, and develop a program to monitor the short- and long-term impact of those decisions.

References

Abedi, J. (2009). Comparability issues in the alternate assessment based on modified achievement standards for students with disabilities. In M. Perie (Ed.), *Considerations for the Alternate Assessment based on Modified Achievement Standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment* (pp. 267-294). Baltimore: Paul Brookes Publishing Co.

Barton, K.E., & Winter, P., (2010). Evaluating the comparability of scores from an alternative format. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 95-104). Washington, DC: Council of Chief State School Officers.

Bennett, R.E. (2003). *Online assessment and the comparability of score meaning* (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Lennon, M.L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191-205.

Cormier, D.C., Altman, J.R., Shyyan, V., & Thurlow, M.L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Davis, L.L., Kong, X., & McBride, M. (2015, April). *Device comparability of tablets and computers for assessment purposes*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Davis, L.L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment, 20*, 180-198.

Davis, L.L., & Strain-Seymour, E. (2013a, June). *Digital devices research*. Paper presented at the CCSSO National Conference on Student Assessment, National Harbor, MD.

Davis, L.L., & Strain-Seymour, E. (2013b). *Keyboard interactions for tablet assessments.* Washington, DC: Pearson Education. Retrieved May 2, 2016, from http://researchnetwork.pearson.com/wp-content/uploads/Keyboard.pdf.

Davis, L.L., Strain-Seymour, E., & Gay, H. (2013). *Testing on tablets: Part II of a series of usability studies on the use of tablets for K-12 assessment programs.* Washington, DC: Pearson Education. Retrieved May 2, 2016, from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-Part-II_formatted.pdf.

Eberhart, T. (2015). *A comparison of multiple-choice and technology-enhanced item types administered on computer versus iPad* (Doctoral dissertation).

Keng, L., Davis, L., McBride, Y., Glaze, R., & Steedle, J. (2015). *Spring 2014 digital devices comparability research study*. Washington, DC: Partnership for Assessment of Readiness for College and Careers (PARCC).

Keng, L., Kong, X.J., & Bleil, B. (2011, April). *Does size matter? A study on the use of netbooks in K-12 assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kieffer, M.J., Lesaux, N.K., Rivera, M., & Francis, D.J. (2009). Accommodations for English language learners on large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79, 1168-1201.

Kingston, N.M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22-37.

Laitusis, C., Buzick, H., Stone, E., Hansen, E., & Hakkinen, E. (2012, June). *Literature review of testing accommodations and accessibility tools for students with disabilities.* Princeton, NJ: Smarter Balanced Assessment Consortia, Educational Testing Service and Measured Progress. Retrieved from: http://www.smarterbalanced.org/wordpress/wp- content/uploads/2012/08/Smarter-Balanced-Students-with-Disabilities-Literature-Review.pdf

Li, H. & Suen, H.K. (2012). Are test accommodations for English language learners fair? *Language Assessment Quarterly, 9*(3), 293-309.

Lorié, W. (2015, March). *Reconceptualizing score comparability in the era of devices.* Presentation at the Association of Test Publishers conference, Palm Springs, CA.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449-458.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30, 10-28.

Pisacreta, D. (2013, June). *Comparison of a test delivered using an iPad versus a laptop computer: Usability study results.* Paper presented at the CCSSO National Conference on Student Assessment (NCSA), National Harbor, MD.

Rogers, C.M., Lazarus, S.S., & Thurlow, M.L. (2014). *A summary of the research on the effects of test accommodations, 2011-2012* (Synthesis Report 94). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Rogers, C.M., Christian, E.M., & Thurlow, M.L. (2012). *A summary of the research on the effects of test accommodations: 2009-2010* (Technical Report 65). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Powers, D.E., & Potenza, M.T. (1996). *Comparability of testing using laptop and desktop computers* (ETS Rep. No. RR-96-15). Princeton, NJ: Educational Testing Service.

Renaissance Learning (2013). *Comparability study: STAR Enterprise iPad and web application versions.* Wisconsin Rapids, Wisconsin: Renaissance Learning, Inc.

Sanchez, C.A., & Branaghan, R.J. (2011). Turning to learn: Screen orientation and reasoning with small devices. *Computers in Human Behavior*, *27*(2), 793-797.

Sanchez, C.A., & Goolsbee, J.Z. (2010). Character size and reading to remember from small displays. *Computers & Education*, *55*(3), 1056-1062.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005).

Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series (NCES 2005–457). Washington, DC: U.S. Government Printing Office. Retrieved May 2, 2016, from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment*, *26*(4), 284-292.

Sireci, S.G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational researcher, 34*(1), 3-12

Strain-Seymour, E., Craft, J., Davis, L.L., & Elbom, J. (2013). *Testing on tablets: Part I of a*

series of usability studies on the use of tablets for K-12 assessment programs (White paper). Washington, DC: Pearson.

Texas Education Agency (2008). *A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests.* Retrieved from http://ritter.tea.state.tx.us/student.assessment/resources/techdigest/Technical_Reports/2008_literature_review_of_comparability_report.pdf

U.S. Department of Education. (2015). Peer Review of State Assessment Systems, Non-Regulatory Guidance for States. September 25, 2015. Washington, DC: USED. Retrieved May 2, 2016, from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement*, Vol 2. Abingdon, UK: Routledge.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olsen, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*(2), 219-238.

Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olsen, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement*, *68*(1), 5-24.

Winter, P. (2010). Comparability and test variations. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 1-11). Washington, DC: Council of Chief State School Officers.

Yu, L., Lorié, W. & Sewall, L. (2014, April). *Testing on tablets*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Appendix: State Action Guide for Gathering Evidence to Support Claims of Comparability across Computerized Devices

Our reality is that states will develop computer-based tests and districts and schools will use different devices to administer those tests. It is also true that all stakeholders will want to use and interpret results from those assessments interchangeably – without concern or regard for the device on which the test was administered. There is no question that the use of different computerized devices is a move away from standardization that poses a threat to score comparability. However, there are clear, practical steps throughout the assessment cycle that states and their assessment contractor(s) can take to be proactive in identifying, anticipating, and avoiding potential threats to score comparability due to devices such as

(a) minimize known threats to score comparability,

(b) document evidence of score comparability, and

(c) monitor potential threats to score comparability.

As a starting point, states must clearly define issues related to the impact of the use of different devices on score comparability and isolate those issues from other threats to comparability.

- *Identify the Comparability Concern(s) Being Addressed*

  The use of different devices is likely to be only one of many threats to score comparability faced by a state administering computer-based tests. For example, establishing the comparability of content across different forms or versions of an assessment will always be a pressing concern that is independent of differences in devices used to administer the assessment. Identify each of the likely threats to score comparability and its relationship to the use of different devices.

- *Determine the Desired Level of Comparability*

  Most states are primarily concerned with reporting student scores on the same score scale and aggregating scale scores across tests administered on different devices with no regard to the device used to administer the assessment. Identify which reported scores are intended to be comparable (e.g., scale scores, performance levels) and which, if any, are not.

- *Clearly Convey the Comparability Claim or Question*

  In most cases, states will likely want to argue that students are taking the state assessment under the conditions most likely to produce the most accurate estimates of their true achievement. Therefore, plan to compile evidence parallel to that which is used to support the use of accommodations (Critical Element 5.3).

- *Focus on the Device*

  Focus on the three areas most directly related to the device: the manner in which content is presented, the manner in which students interact with that content, and the manner in which students respond to that content.

Below we outline three broad steps[7] meant to produce evidence that supports claims of score comparability across devices. The analyses outlined in these steps also have the potential to produce evidence of *in*comparability, thus those investigating the question of comparability should be prepared for such results, and in the presence of such results develop plans to investigate and minimize potential causes of incomparability. Such steps require that the devices used to administer the assessments are known and the devices can, and often do, differ from those approved by the state for testing. Thus states and other agencies examining device comparability may first need to conduct a survey to determine what devices were actually used for the administrations in question.

These steps, when taken together, represent an ongoing approach to quality assurance (QA) and quality control (QC). We envision that the types of analyses outlined below will be incorporated into a state's QA and QC plan, so that investigations of device comparability become routine and that procedures are in place when a new device is introduced. The ordering of the steps reflects the way in which states may want to structure a plan for peer review – states may consider conducting the earlier steps in the immediate future and slate the latter steps for investigation in the upcoming years as data becomes available. In crafting a plan to examine device comparability, states will need to determine what evidence is already available and what will need to be created. States will also need to determine who will be providing the necessary data and who will be conducting the analyses (e.g., the state, their vendor, or a third party contractor). States that are crafting new requests for proposals for their assessment systems may consider including text addressing these types of concerns.

In addition, the steps below are generally ordered so that the most demanding, in terms of data requirements and methodology, analyses are presented last. These steps are intended to be a useful starting point for state-specific planning purposes as the particular steps a state needs to conduct are dependent on the comparability claim(s) it is making, as well as the context of its testing program.

MINIMIZE THREATS TO SCORE COMPARABILITY

The process of minimizing threats to score comparability begins with the design and development of the assessment. As states and their assessment contractors are beginning to design and develop assessments, they must ensure that their processes include steps designed to identify and reduce threats to score comparability across devices. The state's request for proposals (RFP) and the contractor's response should demonstrate an awareness of the leading edge of research on issues that impact score comparability across devices. At the outset, the state and assessment contractor should clearly identify and agree on basic steps that will be included in the design and development process to minimize threats to score comparability. Two such steps are conducting a Functionality Review and planning for the use of Cognitive Laboratories. In addition to those two steps, we recommend that states and vendors apply the known research findings on reducing threats to score comparability and best practices in the field. Each of these four steps is described in more detail below.

7 Way et al. (2016, p. 277-278) provides a table of sample comparability questions and methods that nicely complements these steps.

Functionality Review

This step entails examining that the items are displayed in the same way across all approved devices; for example, ensuring that the item text is not awkwardly broken apart on one device (e.g., when rescaled to fit on small screens), relative position and size of graphics are similar, graphics are not distorted on device, and input options are equally functional. One approach to conducting a functionality review is to simply examine each item in the test administration platform on the devices side by side. This step is, to some degree, a bare minimum. In addition, Way et al. (2016, p. 279) recommend that this type of review be built into the item development process – that item development tools allow item writers to write items then immediately see how they will be displayed across a number of different devices.

Cognitive Laboratories

This step takes the principles outlined in the functionality review one step further by examining how students' cognition differs when the same items are presented on different devices. Widely used in both educational measurement as well as more broadly in the social sciences, cognitive laboratories are a method by which students are asked to "think aloud" while completing a task, e.g., responding to an item, so that their cognition can be examined. In work on score comparability across devices, cognitive laboratories have shown that decreasing the amount of information displayed on screen, e.g., by scrolling (Sanchez & Goolsbee, 2010; Sanchez & Branaghan; 2011) or though pop-up tools like a calculator or on screen keyboard (Davis & Strain-Seymour; 2013b), increases the demands on students' working memory.

Given limited resources, cognitive laboratories can be targeted to new item types, tools, or devices that are being introduced by the state and its contractor which do not have a research base to support their use. If resources permit, areas which have been identified as posing the greatest threats to comparability in the past could also be examined through cognitive laboratories (e.g., open-ended items like responses to writing prompts and subjects like reading, which often contain long passages involving scrolling; writing, which often involve long open-ended responses; or geometry, which involves the physical or mental manipulation of shapes).

Apply Research on Score Comparability across Devices

Although research on the impact of devices on score comparability is still a nascent area of study, there are some common findings that are beginning to emerge that provide guidance for states to following in the design, development, and administration of their state assessments. The following table is created for the purpose of documenting research-based steps states can take to minimize threats to score comparability across computerized devices. It is our intention that this table be only a starting point. As more studies are conducted and as the body of literature grows, we intend for this table to be updated and bolstered to provide a running record of the latest knowledge in this area.

Table 1

Minimizing Threats to Comparability During Test Design, Development, and Administration

Recommendation	Citation
Standardize Content Across Devices The amount of information shown on screen at any one time is constant across devices.	Winter 2010; Bridgeman, Lennon, & Jackenthal, 2003; Sanchez & Branaghan, 2011
Device Familiarity and Fluency Provide students the opportunity to become familiar with and develop fluency on the devices used for assessment. Provide tools to test students on their device fluency to ensure they have the minimum required set of skills (e.g., toggling between alpha and numeric keyboards on a tablet) to access the tested content.	Lorié, 2015
Screen Size Establish parameters for minimum screen size. Current research suggests screens are of 10" or larger reduce threats to score comparability.	Keng, Kong, & Bleil, 2011; Davis, Strain-Seymour, & Gay, 2013
Standardize Embedded Tools Across Devices If it is necessary to allow for on-screen tools that are specific to any one device (e.g., on-screen keyboard), to the extent practicable, do not block or otherwise prevent access to any part of the assessment content.	Davis & Strain-Seymour, 2013b
Touch Screens If touch screens are used, the objects requiring input or interaction are sufficiently large (e.g., bigger in size than students' fingertips) and spread apart as to avoid issues with precision.	Strain-Seymour, Craft, Davis, & Elbom, 2013; Eberhart, 2015
Understand How Technology-Based Tool Are Used During Testing For example, because the use of a mouse allows students to track their reading, it may be beneficial to ensure that additional tracking tools are allowed for students using touchscreens without a mouse.	Way, Davis, Keng & Strain-Seymour, 2016; Eberhart, 2015
Understand the Relationships Between Technology and Specific Tests or Tasks For example, if possible, provide students with external keyboards when responding to open-ended or composition items.	Davis & Strain-Seymour, 2013b; Pisacreta, 2013

Follow Best Practices

In addition to applying research-based practices to the design and development of their assessments, states should also adhere to best practices in the design, development, and administration of large-scale assessments. Table 2 contains a list of steps that states can take throughout the assessment cycle to minimize threats to score comparability across devices. Just as Table 1 is designed to be updated and modified as the research emerges, so too is this table a starting point. We intend for this table to be continuously updated as states submit their own evidence to peer review.

Table 2

Following Best Practices to Minimize Threats to Comparability

Type	Action
Test Design	• Develop item specifications that require that items render the same and are equally functional across all devices and modes of delivery (e.g., touchscreen input vs. mouse input)
List of Approved Devices	• Generate a list of approved devices that have been tested and certified by the test delivery vendor. This list is specific to not only the device, but the approved operating systems and software versions that are supported. • Identify a list of security features that must be included and verified on the approved devices such as internet lockdown and removal of screen reading apps. • Establish protocols for collecting information on the devices that are actually used during testing. • Anticipate that there will be requests/attempts to use unapproved devices during testing and establish a policy and procedures for handling those situations.
Administration Procedures	• Include training materials on device functionality within the administration guides. • Provide opportunities for administrator and student training and practice in the use of devices that will be used to administer the assessment. In particular, identify any tools that will be used during the assessment with which administrators and students may not be familiar. • Establish policies on the use of devices during an administration cycle, including addressing those that may impact student performance (e.g., allowing students to switch devices between subject tests in order to access an external keyboard for essay items).
Plans for Continued Quality Assurance	• Establish a quality assurance (QA) and quality control (QC) plan that identifies processes and procedures to be followed by the state, its assessment contractors, and local test coordinators/administrators, as appropriate. • Identify a set of planned analyses to support score comparability across currently approved devices along with an ambitious yet attainable timeframe. • Establish a long-term plan to continue to evaluate device comparability, including a way to field test new devices before approval for operational use.

Document Evidence of Score Comparability

Evidence to support score comparability can be classified into three categories. The first two categories are related directly to the steps described above to minimize threats to score comparability. In other words, if the steps outlined above are followed, the documentation resulting from those measures to minimize threats to score comparability is what we suggest providing as evidence for the first two categories presented in this section. The third category involves post hoc analyses that can be conducted to determine whether there are differences in student performance that can be attributed to the use of different devices.

Evidence related to test design and development

The first category of evidence is related to all of the activities designed to minimize threats to score comparability across devices implemented by the state and its assessment during the design and development of the assessment. Evidence to support score comparability in this category includes documentation of all relevant decisions and actions taken during assessment design and development including documentation of

- the design and results of the Functionality Review,
- the use of and findings resulting from Cognitive Laboratories, and
- detailed test, item, and device specifications.

Evidence related to test administration

The second category of evidence is related to the use of best practices designed to minimize threats to score comparability in the administration of the assessment. Evidence to support score comparability in this category is directly related to the information described in Table 2. In addition to providing evidence that the identified policies and procedures have been established, states should also be able to document evidence of the implementation and effectiveness of those policies and procedures, as appropriate. Examples of evidence of implementation and effectiveness may include

- A list of the devices actually used by districts and schools during test administration, including identification of any deviations from the list of approved devices or device specifications
- Documentation that appropriate district and school personnel received training in the use of devices for the administration of the assessment
- A report on the use of training and/or practice materials by test administrators and students prior to test administration
- Evidence of the effectiveness of materials and training, including information collected via surveys and focus groups
- Documentation of device-related problems during test administration
- If available, indicators that students have acquired a desired level of familiarity and fluency with the device(s) they will use during testing
- Documentation of a detailed Quality Assurance and Quality Control Plan that has been reviewed by appropriate external reviewers such as the state's Technical Advisory Committee

Evidence related to test performance

The third category of evidence includes evidence related to student performance on the assessment. After students have taken the assessment, the results can be used to examine whether there are differences in performance related to differences in the devices. Approaches to examining the results range from descriptive to causal. These approaches also have a range of data requirements – some

require access to student item responses, others scale scores. Quasi-experimental approaches require access to data on students' past test scores and background characteristics. Approaches examining relationships to external variables like first-year college GPA require access to those external variables. Below we list several approaches that can be taken to compare performance across devices, again arranged in order from those that are straightforward and require minimal effort to those that are more complex, and from those that recommended basic requirements for all programs on an annual basis to those that can conducted on a cyclical basis and/or would be desirable if resources permit.

Checking Item Responses

There are simple checks that can be conducted on student item responses that can be done almost immediately after the item responses are recorded and scored. These types of checks are not meant to replace other examinations of student performance, but to help inform which content areas and grades such examinations (or which devices) should be a focus of investigation. A state may wish to expedite this type of analysis by using the extant data – simply all of the students taking the assessments on each device – or may wish to conduct comparisons among matched groups of students so that differences in performance are more directly attributable to differences in devices. After the tests are scored, item difficulties (e.g., proportion correct or p-values) can be compared across devices – with interpretations supported by relevant information on the samples of schools or students using each device. In addition, the length of time spent on each item (or the test overall), as well as the length of responses to open ended items, can be compared. Depending on how the test administration platform and related data management systems capture and store response data, accessing data related to the time spent on each item and the length of open ended responses may be straightforward, or it could be so time consuming that the state may choose to focus their efforts elsewhere.

Examining for Differential Item Functioning

In addition to examining p-values and response length/time, states should consider conducting differential item functioning (DIF) analyses to detect systematic differences in responses across the different computerized delivery devices. While there are a number of analytic techniques for detecting DIF, states will need to be aware of the information each provides and determine how to set thresholds for flagging potentially problematic items. When non-uniform DIF is present, states may want to consider detecting DIF for the overall assessments (e.g., comparing test characteristic curves) rather than at the item level. When assessments contain clusters of related items, states may want to consider approaches to detecting DIF for those complete clusters as well as for individual items.

Comparing Test Scores

Once produced, students' scale scores can be used to examine differences in performance across devices through comparisons to historical performance or comparisons across contemporaneous groups of students. In terms of the former, the performance of groups of students who switched devices between years provides a way to examine the influence of device differences on performance. For example, finding that the scale scores produced by students testing on tablets, but in the prior year tested on desktop computers, are in line with their expected scale scores (e.g., as quantified through a regression based

approach) would be evidence to support score comparability between the tablets in question and desktop computers. In terms of the latter, the performance of matched groups of students taking the assessment on different devices can be compared using methods like coarsened exact matching or propensity score matching. Such matching approaches have been widely used in other areas of comparability.

Internal Structure

States may want to consider comparing the factor structures of the assessments delivered across different computerized devices in order to gather evidence of measurement invariance. Evidence of measurement invariance would support claims of comparability because it means that the assessed construct is measured in the same way, regardless of the delivery device. As with the other analyses suggested in this section, there are a range of analytic techniques that could be employed to approach the question of measurement invariance. These techniques range in complexity from comparing estimates of internal consistency (i.e., Cronbach's a reliability coefficient), to using advanced modeling methods such as confirmatory factor analysis and structural equation modeling to test for device-specific factors that may be contributing to construct-irrelevant variance in the score estimates generated from different delivery devices.

Relationships between Test Scores and External Variables

Comparability evidence can additionally come from criterion-related or predictive validity evidence. If scores produced across different computerized devices are truly measuring the same construct, it would be expected that the scores have equivalent correlations with external variables. To examine this for the purpose of gathering evidence of comparability, states would correlate test scores resulting from different devices with easily accessed criterion variables of interest (e.g., previous year achievement, GPA, other assessment scores). The strength of the relationships between the test scores produced from different devices and the criterion variables should be the same – or within the bounds of sampling error – in order to lend support to claims of comparability in score interpretations.

MONITOR POTENTIAL THREATS TO SCORE COMPARABILITY

For both commonly used and newly developed devices, the state should collect information on the devices used and develop a plan to monitor performance across devices on a regular basis. In accordance with best practices, even in cases where one expects there to be no threat to score comparability (e.g., field test items are embedded in test forms using a matrix-sampled approach, items are administered to students in a random or non-uniform manner, or accommodations are allowed), the state should conduct secondary analyses periodically to show that there is no negative impact on score comparability. As with the use of accommodations, such a monitoring program is particularly important when there is local control over the manner in which policies are implemented. With accommodations, states regularly monitor local policies and actual accommodations use to ensure that allowable accommodations are not being overused or underused by some school districts. Similarly, with the use of technology-based devices, the state should be aware of any systematic differences in the use of devices across districts and be prepared to investigate the potential impact of any such systematic differences that might be found.

Council of Chief State School Officers

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