Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

The Comparability of Scores from Different Digital Devices: A Literature Review and Synthesis with Recommendations for Practice

Nathan Dadey, Susan Lyons & Charles DePascale

To cite this article: Nathan Dadey, Susan Lyons & Charles DePascale (2018) The Comparability of Scores from Different Digital Devices: A Literature Review and Synthesis with Recommendations for Practice, Applied Measurement in Education, 31:1, 30-50, DOI: 10.1080/08957347.2017.1391262

To link to this article: https://doi.org/10.1080/08957347.2017.1391262

Published online: 02 Jan 2018.

APPLIED MEASUREMENT IN EDUCATION
2018, VOL. 31, NO. 1, 30–50
https://doi.org/10.1080/08957347.2017.1391262

The Comparability of Scores from Different Digital Devices: A Literature Review and Synthesis with Recommendations for Practice

Nathan Dadey, Susan Lyons, and Charles DePascale

The National Center for the Improvement of Educational Assessment, Inc.

ABSTRACT
Evidence of comparability is generally needed whenever there are variations in the conditions of an assessment administration, including variations introduced by the administration of an assessment on multiple digital devices (e.g., tablet, laptop, desktop). This article is meant to provide a comprehensive examination of issues relevant to the comparability of scores across devices, and as such provide a starting point in designing and implementing a research agenda to support the comparability of any assessment program. This work starts with a conceptual framework rooted in the idea of a comparability claim—a conceptual statement about how each student is expected to perform on each of the devices in question. Then a review of the available literature is provided, focusing on how aspects of the devices (touch screens, keyboards, screen size, and displayed content) and aspects of the assessments (content area and item type) relate to student performance and preference. Building on this literature, recommendations to minimize threats to comparability are then provided. The article concludes with ways to gather evidence to support claims of comparability.

Introduction

Computer-based assessment introduces new threats to the comparability of assessment results, including those introduced by differences in technological devices. As assessments move away from pencil and paper, 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, large-scale, computer-based assessments are administered to students via a wide variety of 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 in the use of assessment results for high-stakes purposes, test developers and test users must be able to confidently claim that results are not impacted by variations introduced through the use of different devices.

Evidence of comparability is needed when there are variations in the conditions of an assessment administration (cf., *Standards for Educational and Psychological Testing, AERA, APA & NCME*, 2014). Evidence is needed of comparability between pencil-and-paper and computer-based administrations; as well as among the different devices used to deliver a computer-based assessment (e.g., tablet, laptop, desktop). This article provides context and considerations for operational testing

CONTACT Nathan Dadey ✉ ndadey@nciea.org The National Center for the Improvement of Educational Assessment, Inc., 31 Mount Vernon St., Dover, NH 03820.

programs, as well as for an overall research agenda, in addressing the comparability of scores produced by students taking the *same assessment* on different devices—often referred to as device comparability. This term, however, does not imply that the devices are comparable, rather that the scores resulting from assessment administrations on different devices are comparable.

This article is akin to the work of Way, Davis, Keng, and Strain-Seymour (2016) in that it provides a broad overview of issues related to score comparability across devices. However, the work of Way et al. (2016) establishes key areas of concern in broad strokes, whereas this study is meant to provide a detailed summary and synthesis of empirical examinations of student performance and preference across devices. That is, this work provides an updated review and synthesis of the literature, focused on how student performance differs across devices and their features, as well as how student familiarity relates to performance. This work also provides recommendations on ways to minimize potential threats to score comparability and gather evidence to support claims of comparability.

In the first section, we provide a theoretical framework for discussing the comparability of scores. We define comparability and outline two example claims that can be made about the comparability of assessment scores. The second section contains a summary of the existing research dealing directly with the comparability of scores across various devices. This summary is detailed—the reader may be well served by using the synthesis of the third section to guide their reading of the second. Also, the structure of the summary is meant to reflect the evolution of research on comparability across devices. The third section contains a synthesis and interpretation of that research. The synthesis includes our recommendations for minimizing threats to comparability as well as recommendations of analyses that test developers and users could employ to support comparability claims.

Theoretical Framework

Defining Comparability

The body of literature on score comparability is expansive and varied, as are the definitions of score comparability itself. At times the methods used to produce evidence of score comparability are equated with comparability—for example, if the assessment data does not display differential item functioning across the test variations, then the scores are comparable. We take the view that claims of score comparability should be supported by evidence produced by multiple methods; and thus, score comparability is not defined or determined by a single method or analysis. This view is compatible with the often cited definition of comparability as interchangeability—“when comparability exists, scores from different testing conditions can be used interchangeably” (Bennett, 2003, p. 2; see also Winter, 2010; Way et al., 2016). Building on this definition, Winter (2010) conceptualized comparability as requiring that a “test and its variations must [1] measure the same set of knowledge and skills at the same level of content-related complexity (i.e., constructs); [2] produce scores at the desired level (i.e., type) of specificity that reflect the same degree of achievement on those constructs; and [3] have similar technical properties (e.g., reliability, decision consistency, subscore relationships) in relation to the level of score reported” (p. 3). We extend this specificity by suggesting that test developers and users must clearly articulate their intended *claim* of comparability, and do so before they collect evidence. Defining a claim allows developers and users to lay out *a priori* evaluative criteria about what constitutes enough evidence.

The Comparability Claim

The definition of comparability provides a starting point for understanding the necessary evidence to support score comparability, but alone is insufficient for informing investigations of comparability. We suggest that the underlying statement, or claim, being made about the comparability of student performance be explicitly articulated. Different claims of comparability will require either the collection of different evidence or different analysis and interpretation of the same evidence. Consider the following two example claims:

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

The first claim is that student scores are device-neutral. In contrast, the premise of the second is that a student may 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 claim places a premium on the traditional concept of standardization. Everyone takes the assessment under the same conditions and, therefore, the intended users of the results know the conditions under which the test was administered. For example, those assessments used for college admissions have traditionally required administration conditions that support this claim. The second statement reflects an increased acceptance of flexibility, similar to current practices around individual student accommodations within most assessment programs. Test accommodations are offered under the assumption that they allow for improved access to the intended construct. For example, most students with Individualized Education Plans (IEPs) are encouraged to take the assessment under conditions that most closely mirror instruction. As long as that flexibility is determined to not alter the construct, 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, we believe that most operational testing programs are likely to make a comparability claim that allows for flexibility, rather than one that argues that scores are device-neutral. That is, test developers and users are more likely to claim that allowing for choice in the test administration device removes barriers to performance (e.g., unfamiliarity). In this way, devices will not be regarded as interchangeable as if they were No. 2 pencils. We have arrived at this conclusion, in no small part, because of the diversity of devices used in classroom instruction and our suspicion that one type of device is often used predominately. Although we argue that many assessment programs may favor this claim, increased emphasis on flexibility is *not* coupled with a decreased need for comparability and evidence to support it. Any claim an assessment program might make requires supporting evidence. The differences are in what evidence is required to support judgments of score comparability. Claims rooted in flexibility would likely need to produce evidence akin to those outlined within the fairness chapter of the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014)—which ask that developers and users articulate what is, and is not, construct irrelevant variance and how construct irrelevant variance is reduced through non-standardization.

In the next section, we review the literature dealing with the comparability of scores across devices. We summarize student performance and preference without explicitly tying the results to either claim. Much of the literature has focused on detecting differences in performances; differences, that if found, could refute claims based on standardization but could support claims based on flexibility. However, such support would only be garnered if differences in performance corresponded to differences in familiarity or fluency—most concretely reflected in the devices used for instruction. In addition, investigations of comparability will require the collection of evidence above and beyond that afforded in the existing literature. Our view is that any comparability investigation must first explicitly articulate the comparability claim, then determine what evidence is needed and how that evidence will be evaluated, and only then collect and review the evidence to form a coherent and convincing justification for the stated comparability claim.

A Review of Student Performance and Preference Across Devices and Their Features

The number of educational assessments administered electronically has expanded dramatically in recent years, as has the number of different devices used to administer them. Despite widespread use of electronic devices for test delivery, studies on the comparability of scores across different devices are still relatively rare. Due to the limited body of research, we have drawn not only on studies published in academic journals or presented at conferences, but also from gray literature[1]. To gather this unpublished literature, we contacted 17 testing organizations. All of these organizations responded and eight were able to provide reports of completed studies that specifically examine the comparability of scores resulting from the use of different electronic devices. Those studies, along with studies uncovered via traditional literature review of academic journal and conferences are included in this review.

In the review we first summarize study results that compare student performance and preference on one device to another device (e.g., computer vs. tablet) and also examine the role student familiarity plays. We then summarize results that compare student performance and preference in relation to specific device features (e.g., screen size, keyboard). This latter set of study results reflects an emerging focus on the specific factors that influence how students interact with, and thus perform on, a device. When needed, we draw on a single study multiple times. The Appendix contains a summary table of the studies referenced.

Effects Across Devices

Performance

Some of the earliest studies examined differences in student performance, also referred to as "device effects" in a fashion analogous to mode effects, and student preferences between laptop and desktop computers. In 1996, Powers and Potenza found no differences in Graduate Record Examination (GRE) Verbal and Quantitative scores between laptop and desktops, and small differences in GRE Writing scores in favor of desktops. Participants reported difficulties with the size and layout of the laptop keyboard, difficulties they did not have with the desktop keyboard. At the time of this study, however, participants reported little prior experience with laptop computers, thus these findings may not generalize. More recently, a 2002 experimental study of the National Assessment of Educational Progress's (NAEP) Technology-Based Assessments found that student writing performance was slightly lower for students using laptop computers than for students using desktop computers (Horkay, Bennett, Allen, & Kaplan, 2005, see pp. 24–31). However, this difference in performance only occurred on one of two essays in a small sample of randomly assigned students. Using a larger, non-random sample, the authors came to slight different conclusions—female students performed significantly lower on both essays when using NAEP-provided laptops, relative to school desktops. This difference was not found for male students and in the combined sample the differences in performance were not significant.

Current work has shifted away from comparisons between desktop and laptop computers. Instead, comparisons are now generally between computers—both desktop and laptops—and tablets. In a comparison between high school students randomly assigned to test on a computer or tablet (full sized, 9.7" iPad), Davis, Kong, and McBride (2015) found no statistically significant difference in mean scale scores on a multi-section assessment of high school reading, math, and science. In a subsequent study using the same dataset, Davis, Morrison, McBride and Kong (2016) compared performance between computers and tablets at each score point (instead of the scale score averages used in the prior study) for all students and by gender and ethnicity subgroups. They found no significant differences for math and science across the scale score distributions, both overall and by

[1] Gray literature is a term used to refer to information found outside of traditional academic publications. Among other types of information, it includes unpublished manuscripts, technical reports, and white papers.

subgroup. They did, however, find that male students performed better under the tablet condition on the reading assessment, resulting in differences at lower end of the scale score distributions. Davis, Orr, Kong, and Lin (2015) also found no statistically significant differences on essay scores from students randomly assigned to laptop computers, 10” tablets with a virtual keyboard, or 10” tablets with a physical keyboard. Notably, the sample of students in these two studies had more experience with some devices (Davis, Orr, et al., 2015; Davis, Morrison, et al., 2016), but these differences did not translate into differences in performance.

In one of the largest studies to date, conducted by Davis, Janiszewska, Schwartz, and Holland (2016), compared students’ performance on computers (laptops and desktops) to tablets (with and without external keyboards), on Australia’s National Assessment Program in Literacy and Numeracy in years 3, 5, 7, and 9. In both reading and numeracy, they found no statistically significant differences in student performance in years 3 and 5 across conditions. However, they did find that students in years 7 and 9 performed slightly better on computers and tablets with external keyboards, relative to tablets without external keyboards. Eberhart (2015) came to similar conclusions, finding that 7th grade students performed slightly better when taking 7th grade English Language Arts (ELA) and mathematics assessments on a computer, relative to a tablet.

In a study involving very different devices, Schroeders and Wilhelm (2010) compared scores from assessments administered on small, handheld Personal Digital Assistant (PDA) devices (3.7 inch display) with scores from assessments delivered on laptops. From their sample of German high school students, they found that students generally answered more items correctly on laptops relative to PDAs, but the effect was modest. Using confirmatory factor analysis, Schroeders and Wilhelm (2010) also found small device-specific factors for the PDA and the laptop. They hypothesized that the factors may be attributed to differences in familiarity with device, item presentation modes, or differing motor skill and perceptual demands—suggesting that further research would need to investigate these factors. Yu, Lorié, and Sewall (2014) work is one such investigation. The authors conducted cognitive laboratories and found that though few students experienced frustration, more students indicated frustration when responding to items on a tablet than on a laptop. In the same study, a higher number of students indicated that they would rather take a high-stakes exam on a computer than on a tablet. Such preferences suggest that familiarity might play a role in how students interact with a device.

Familiarly

Student familiarity with a device and its features (e.g., external keyboard, touch screen) is frequently cited as key in considerations of comparability (see Davis, Janiszewska, et al., 2016; Keng, Davis, McBride, Glaze, & Steedle, 2015; Lorié, 2015; Powers & Potenza, 1996; Horkay et al., 2005; Schroeders & Wilhelm, 2010). Lorié (2015) argued 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 to ensure they are able to access the tested content.

Davis and colleagues (2016) examined device familiarity and fluency through their study design, in which they crossed the device used to assess students (computer, tablet, and tablet with an external keyboard) with the device used for instruction within each school (again, computer, tablet, and tablet with an external keyboard). They found some quantitative evidence that students familiar with tablets with external keyboards through classroom use perform better when tested on these devices, particularly in later grades. Davis and Strain-Seymour (2013a) found that familiarity with devices varies by age, and that device preference may also vary by content. For example, 11th graders reported using tablets in school more frequently than laptops, while 5th 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 device fluencies. Device fluencies may also encompass navigation—for example, students have reported difficulty with scrolling on iPads, as the scrollbar is not visible by default, instead students must swipe up or down, causing the scroll bar to appear (Pisacreta, 2013). The effects of device features that often require device fluency, like screen navigation, are discussed next.

Effects of Device Features

The differences in student performance presented in the prior section are not likely due to the device itself, but rather due to certain features of the device. Assessment programs may be able to mitigate the observed effects across devices by better understanding features that may contribute to score differences. This section includes literature that identifies three features of devices—touch screens, keyboards, and screen size/displayed content—and two assessment features—content area and item type—that may be contribute to differences in student performance across devices.

Touch Screens

In general, some input precision is compromised when using a fingertip on a touchscreen instead of a mouse (Way et al., 2016). This reduced precision can cause students to take slightly longer to answer any given item when using a tablet (about 3–4 seconds more; Kong, Davis, McBride, & Morrison, 2016). Generally, students have problems with fingertip input when objects on the screen requiring interaction (e.g., selection, drag-and-drop) are close in size or smaller than the students' fingertips, or when objects are close together (Davis, Janiszewska, et al., 2016; Eberhart, 2015; Strain-Seymour, Craft, Davis, & Elbom, 2013). Additionally, an added benefit of a mouse is the "hovering" function that visually shows the cursor's location on the screen, which can support accuracy of input and guided reading (Eberhart, 2015; Way et al., 2016). Hovering also allows for the use of "tooltip" or "roll over" icons, which have been used to provide students with additional context related to the assessment environment—for example, to indicate input areas (Davis & Strain-Seymour, 2013a; Strain-Seymour et al., 2013).

Keyboard

Studies that examine student performance with onscreen keyboards, as opposed to external keyboards, find that they work equally well for items requiring short written responses, but student responses tend to be reduced in length when using onscreen keyboards for responding to items requiring longer written responses, likely due to fatigue (Davis, Orr et al., 2015; Davis & Strain-Seymour, 2013b; Pisacreta, 2013). Because students cannot rest their fingers atop the onscreen keyboard, they cannot rely on their keyboarding skills and instead typically defer to a "hunt-and-peck" method. This method is generally less accurate and slower than traditional keyboarding. These issues likely explain Pisacreta's (2013) finding that most participants prefer using an external keyboard when responding to essay prompts. However, this preference may not be completely generalizable—Strain-Seymour et al. (2013) found that younger students, who are less experienced typists, do not prefer external keyboards.

Student fluency with the onscreen keyboard is likely key—including knowing how to switch between letters and numbers, and highlighting and moving text. Additionally, because the keyboard often uses valuable screen space that can block test content, students must know how to toggle the keyboard (Davis, Strain-Seymour, & Gay, 2013; Pisacreta, 2013; Strain-Seymour et al., 2013). Davis, Janiszewska et al. (2016, p. 49) found that students who tested on a device used within their classrooms had fewer problems with a number of features, including the onscreen keyboard. Interestingly, this study also found that students who used a tablet with an external keyboard in the classroom performed worse when tested on a tablet alone, relative to those who tested on a computer or tablet with an onscreen keyboard.

Screen Size and Displayed Content

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. Of the two, the second appears to be more important. Results suggest that when the information shown on the screen is held *constant*, screens of 10 inches or larger are suitable for viewing and interacting with assessments, with little evidence of differences in overall test performance or item-level differences (Davis et al., 2013; Keng, Kong, & Bleil, 2011). Screens smaller than 10" may result in differences in student performance—with lower scores on the device with the smaller screen (Davis et al., 2013; Schroeders & Wilhelm, 2010).

However, the amount of information is often not held constant across devices. In these cases, the content displayed is a function of multiple factors, including the screen resolution, font size, and test administration platform. Differences in the amount of content displayed without the need to scroll or page can lead to differences in test performance, particularly for assessments with reading passages. Bridgeman, Lennon, and Jackenthal (2003) found that while mathematics performance remained stable across different computer monitor sizes and resolutions, 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. They suggest that this result is due to the increase in scrolling under low resolution conditions (lower resolutions generally present less information on the screen at once). Chen and Perie (2016) findings could support similar conclusions, as they found significant score differences in 4th grade ELA between students tested on Chromebooks with 14" screens and computers with much larger, higher resolution screens.

One possible explanation for the detrimental impact of insufficient screen space is that 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 (2013a) found that features of the assessment (e.g., calculator tool or on-screen keyboard) can block parts of the test content, potentially adding additional strain to students' working memory. Similarly, Davis, Janiszewska, et al. (2016) found that when students were unable to view items and a reading passage simultaneously, they reported difficulty keeping an item "in their head while reading the passage" (p. 35). Sanchez and Branaghan's (2011) findings support these conclusions, showing that small screen sizes that generally require additional scrolling, can reduce ability in complex reasoning. However, these researchers found that reorienting a small screen from portrait to landscape effectively mitigated the negative effect of the small screen size and that this change seemed to be especially beneficial for lower ability participants.

Content Area

We focus on assessments of mathematics, reading (or English language arts), and writing. Overall, there is evidence of differential performance across devices in each content area, and these differences tend to appear more often in later grades (e.g., grades 7 to 12, with grade 4 ELA being a notable exception) and generally favor computer-based testing conditions. Potential reasons for these differences included familiarity and fluency (Davis, Janiszewska, et al., 2016), difficulty with touch screen input (Eberhart, 2015) and differences in the amount of displayed content (Bridgeman et al., 2003; in the context of a comparison between desktop computers and laptops). However, only a few studies investigated the causes of differences in performance and even then, the suggested factors do not fully explain the *pattern* of differences (i.e., why differences appear more often in later grades across the three subjects).

In terms of math, several studies have found statistically significant differences, favoring computers, in student scores between computers and tablets in particular grades (Davis, Janiszewska, et al., 2016; Eberhart, 2015; Renaissance Learning, 2013). Renaissance Learning (2013) examined math scores on the STAR assessments in grades 1 to 10 and found statistically significant differences in grades 1 to 3, 5, 7, and 8. Davis, Janiszewska, et al. (2016) investigated grades 3, 5, 7, and 9 on the Australian NAPLAN numeracy assessment and found statistically significant differences in grades 7 and 9. Eberhart investigated student scores on the 7th grade Smarter Balanced assessment also found statistically significant differences. These studies are not entirely consistent with one another, however. For example Davis, Janiszewska, et al. (2016) did not find differences in earlier grades while Renaissance Learning (2013) did. In addition, other studies have not found significant score differences in grades 4, 8, and 10 (Keng et al., 2015), across the entire 4 to 10 grade span (Chen & Perie, 2016), or in high school (Davis, Janiszewska, et al., 2015; Davis, Morrison, et al., 2016). These findings suggest that the samples of students used, the particular tests involved (e.g., content covered or item types), or both, interact with devices in different ways. Blanket statements about the grades in which student performance in math is likely to differ across device are therefore premature.

The findings on reading (or ELA) performance are similar, showing that in some grades, on some tests, students tend to perform slightly better when tested on computers. In an investigation of the STAR reading assessments in grades 1 to 10, Renaissance Learning (2013) found significant differences in grades 1, 5, and 8, favoring the computer condition. Interestingly, the mean differences in grades with nonsignificant results favored testing on tablets. Renaissance Learning (2013) also investigated the STAR Early Literacy assessments in kindergarten to grade 2 and found no significant differences in student scale scores. Chen and Perie (2016), in an investigation of Smarter Balanced ELA assessments in grades 4 to 10, found significant differences for grade 4. Keng et al. (2015) investigation of Partnership for Assessment of Readiness for College and Careers (PARCC) ELA assessments in grades 4, 8, and 10 also found significant differences for grade 4. In contrast, Davis, Janiszewska, et al. (2016) found significant differences in grade 7 and 9, but not in grades 3 and 5. In her examination of the 7th grade Smarter Balanced ELA assessment, Eberhart (2015) found statistically significant differences on one of three forms. Conversely, Davis et al. (2015) found no significant differences in scale scores across computers and tablets on 7th to 10th grade reading items. Davis, Morrison, et al. (2016) re-examined this reading data and found that there were differences in the distributions of scores at the lower end of performance, with those testing on tablets scoring higher. This result was driven by male students who performed better when testing on tablets.

The evidence on writing performance, as captured through extended writing prompts, is more limited than that for math or reading. Three studies have specifically examined differences in performance across devices on writing prompts (Davis, Orr, et al., 2015; Horkay et al., 2005; Powers & Potenza, 1996), but only one has examined student performance on both computers and tablets (Davis, Orr, et al., 2015). In this study, Davis, Orr, et al. (2015) compared mean rubric scores from 5th, 10th, and 11th grade students responding to essay prompts on a laptop, a 10” tablet using the onscreen keyboard, and a 10” tablet using an external keyboard. There were no significant differences in mean rubric scores across the three conditions. This stands in contrast to the two earlier studies (Horkay et al., 2005; Powers & Potenza, 1996) that examined writing on desktops and laptops and found significant differences on rubric scores. Both of these earlier studies examined two writing prompts and found significant differences only for one writing prompt. In addition, these studies are relatively old—students today are likely much more familiar with multiple devices than those studied by Powers and Potenza (1996) and Horkay et al. (2005).

Item Type

Research indicates that student performance across devices varies by the types of assessment tasks (Davis & Strain-Seymour, 2013a; Davis et al., 2013; Eberhart, 2015). 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. Although performance was higher for multiple-choice items on the computer, the same effect was not detected for technology-enhanced items (Eberhart, 2015).

However, Davis, Janiszewska, et al. (2016) and Davis et al. (2015) did not find such effects for multiple choice items. Davis, Janiszewska, et al. (2016), in particular, noted that “using different devices had little to no impact on student performance for item types requiring click interactions (multiple choice, multiple select, and hot spot)” (p. 21) in both reading and numeracy. They did find differences in performance, however, for drag and drop items and text entry items. Drag-and-drop items were easier for students testing on tablets in earlier years (i.e., 3 and 5), but were easier for students testing on computers in latter grades (i.e., 5 and 9). The "think-alouds" conducted by Davis et al. (2013) provide some insight in these results. For these types of drag-and-drop items, students had difficulties dropping objects when the "target area" was too small are too close to other target areas. However, students also reported liking taking drag-and-drop items on the tablet—preferring to directly interact with the items by using their fingertips to drag and drop objects.

Davis, Janiszewska, et al. (2016) found that text entry items, in contrast, were uniformly more difficult on tablets relative to computers or tablets with keyboards. Again, the work by Davis et al. (2013) provides insight—they found the onscreen keyboard often covered the text entry boxes, making it difficult for students to see what they were typing. In addition, Davis et al. (2013) and Pisacreta (2013) both found that typing using an on-screen keyboard is slower and less accurate than using an external keyboard.

Davis et al. (2015) examined seven different item types, including drag-and-drop items, in reading, science, and math. They did not find significant differences in student performance on multiple-choice, drag-and-drop, hot spot, fill in the blank, inline choice, and graph point items. They did, however, find significant difference for multiple-select items, with students performing better on tablets than computers. However, the number of multiple-select items considered was limited, with just six multiple-select items across the three content areas.

Supporting the Claim of Comparability

Computer-based testing will continue to increase, as will the number and types of devices used to deliver those assessments. Inevitability, stakeholders will want to use and interpret results from assessments administered across devices 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, which in turn poses a threat to score comparability. Below we suggest ways that such threats can be addressed as well as how evidence supporting a claim can be gathered. These suggestions are not, however, articulated in terms of a specific comparability claim—assessment developers and users will need to determine the applicability of each suggestion to their own context.

Minimize Threats to Score Comparability

Although the process of minimizing threats to score comparability logically begins with the design and development of the assessment (cf., Way et al., 2016), we suggest steps that can be undertaken at any point in the assessment cycle. These steps include (1) running a functionality review, (2) conducting cognitive laboratories, (3) drawing on current comparability research, and (4) extending current practices for test development and administration.

Functionality Review

This step entails examining whether 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, and input options are equally functional. Doing so could be as simple as examining each item in the test administration platform on the devices side by side. This step is, to some degree, a bare minimum. 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 each item 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. In work on score comparability across devices, cognitive laboratories have shown that decreasing the amount of information displayed on screen, increases the demands on students' working memory (Davis & Strain-Seymour, 2013a; Sanchez & Branaghan, 2011; Sanchez & Goolsbee, 2010). Given limited resources, cognitive laboratories can be targeted to new item types, tools, or new devices, as well as in areas that have been shown be potentially problematic (e.g., open-ended items like responses to writing prompts and subjects like writing, which often involve long open-ended responses).

Draw on Current Research

Although research on the impact of devices on score comparability is still a nascent area of study, there are some emerging findings that provide guidance for test users and developers.

Displayed Content. Keep the amount of information displayed on the screen at any given time constant across devices (Bridgeman et al., 2003; Davis, Janiszewska, et al., 2016; Sanchez & Branaghan, 2011; Sanchez & Goolsbee, 2010; Winter, 2010). Doing so may require adjusting the amount of information displayed on all devices. If on-screen tools 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).

Device Familiarity and Fluency. Provide students the opportunity to become familiar with, and develop fluency on, the devices used for assessment (Lorié, 2015). Provide tools to test students on their device fluencies to ensure they can access the tested content.

Screen Size. Establish parameters for minimum screen size. Current research suggests screens of 10" or larger are sufficient (Davis et al., 2013; Keng et al., 2011).

Touchscreens. Ensure that objects requiring input or interaction are sufficiently large (e.g., bigger in size than students' fingertips) and spread apart to avoid issues with precision (Davis, Janiszewska, et al., 2016; Eberhart, 2015; Strain-Seymour et al., 2013).

Technology-Based Tools. Examine how students use different device features during testing, and if needed, provide similar features across devices. For example, students may use a mouse to track their reading on a computer, and providing similar options for tablets may help support a comparability claim (Eberhart, 2015; Way et al., 2016).

Interactions Between Device Features and Specific Tests or Tasks. Some item types (e.g., drag and drop, text entry, multiple select items) have been shown to be differentially difficult across devices (Davis et al., 2015; Davis, Janiszewska, et al., 2016; Davis & Strain-Seymour, 2013b; Davis et al., 2013; Pisacreta, 2013). For such items, or assessments that contain many items of these types, consider providing students with devices or device features that address potential causes of these differences. For example, students could be provided external keyboards when responding to open-ended or composition items.

Follow Best Practices

In addition to applying research-based practices to the design and development of assessments, test developers and users should also extend current best practices in the design, development, and administration of large-scale assessments. The non-exhaustive list below does so, which we created based our understanding of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on

Measurement in Education, 2014). One point worth emphasizing in relation to this list is that test developers and users may want to assess students on the device most commonly used in instruction. Doing so requires determining what level the matching of students to devices should be conducted (i.e., the classroom, school, or district level), as well as actually determining what devices are used. Doing so also requires the ability to ensure students can take the assessments on the appropriate devices.

Item Specifications. Develop item specifications that require that items render the same, and function equally, across devices and input modes.

Approved Devices. Generate a list of approved devices, as well as approved operating systems, software versions, and peripheral hardware. Establish a policy and related procedures to deal with requests and attempts to use unapproved devices.

Security Features. List security features that must be included and verified on the approved devices, such as Internet lockdown and removal of screen reading apps.

Used Devices. Establish protocols for collecting information on the devices that are actually used during testing, and, if warranted, establish policies to test students on devices they generally receive instruction on.

Training. Include training materials on device functionality within the administration guides. Provide opportunities for administrator and student training and practice on the devices, with particular focus on features that may be unfamiliar.

Variations in Used Devices. Establish policies on variations in device use throughout the administration cycle (e.g., allowing students to switch devices between subject tests in order to access an external keyboard for essay items, providing tablets with cases or stands that allow them to be propped up).

Device Problems. Provide procedures or polices on potential problems with devices (e.g., procedures if a wirelessly connected external keyboard disconnects, if a device loses connection to the Internet).

Document Evidence of Score Comparability

The above section details ways in which comparability can be addressed through the design and administration of an assessment. Supporting a claim of comparability, however, requires that test developers and users also collect and assemble evidence. Some of this evidence comes directly from the actions taken to address issues of score comparability during the design and administration of an assessment. In particular, test developers and users should be able to document the implementation and effectiveness of the policies and procedures implemented as part of the design and administration process. Another key type of evidence, which we turn to next, stems from post-hoc analyses of student performance.

Evidence based on analyses of student performance is also critical to any comparability claim —the body of evidence supporting a claim would be incomplete without some consideration of student performance. The particular analyses needed to support a claim of score comparability, however, are context dependent and therefore often not clear cut. More evidence provides greater support, but those investigating score comparability will need to determine what types of evidence, as well as the amount of evidence, best supports their particular claim. In addition, the results of an analysis could support one claim of comparability but not others. For example, finding that students performed better on tablets, relative to computers, would not support a claim of comparability—if that claim is, say, rooted in the idea that a student's scores should be the same regardless of the device the assessment is taken on. Those same results, however, could support a claim of comparability about matching students to the device that is most commonly used in instruction.

Below we present five types of analyses that can be used to investigate whether student performance differs across devices: examining item responses descriptively, conducting differential item functioning analyses, investigative internal structure, comparing historical or contemporaneous test scores, and relating test scores to external variables. Investigations of test scores, as in the last two analyses, have been and will continue to be important as inferences are made about students upon the basis of these scores. This is not to say that the other analyses are less important—each can offer unique insight into student performance (as shown in Keng et al., 2015). We present the five analyses below in order of their focus— building out from internal checks on items to internal checks on total scores to external checks. Our motivation in presenting these analyses is to demonstrate that: claims of comparability are not tied to the results of any one analysis, multiple analyses should be used to support any claim, and analyses can be built into the assessment cycle at multiple points.

Checking Item Responses

Descriptive statistics of student performance on items can provide a quick check on whether student performance[2] differs across devices. In particular, item *p*-values can provide a way to quickly examine whether items are of similar difficulty across devices. Other item statistics like response times and the length of responses to open-ended items can also provide checks of differences across devices, as can test-level statistics like classical test theory total test scores. These simple checks can be done almost immediately after the item responses are scored, and, potentially, at key points during the administration process. Such checks are not meant to replace other examinations, but to help inform additional investigation.

Examining for Differential Item Functioning

In addition to examining statistics like item *p*-values, test developers should consider conducting differential item functioning (DIF) analyses to detect systematic differences in responses across the different computerized delivery devices. Prior studies (e.g., Chen & Perie, 2016) have found that most items do not display severe DIF. Knowing which items, if any, display DIF allow test developers and users to make informed decisions around the items used to create student scale scores as well as to inform future item development.

Internal Structure

The relationships between student item responses can also provide evidence relevant to claims of comparability. Measures of internal consistency (e.g., Cronbach's α, marginal reliability) can provide an indication about whether scores produced from each device have similar levels of measurement error. In addition to examining internal consistency, test developers and users may also consider investigating measurement invariance[3] using factor analysis or structural equation modeling.

Comparing Test Scores

Once produced, students' scale scores can be compared to historical performance or compared across contemporaneous groups of students. In terms of historical comparisons, the performance of students who switched devices between years provides one way to quantify the influence of device

[2] These descriptive checks should be conducted with a clear idea of how different devices were provided to students. In particular, if devices are not randomly assigned, analysis of item responses may reflect differences in the groups of students, such as prior academic performance, rather than differences in interactions with the devices.

[3] DIF analysis also addresses the issue of measurement invariance.

differences on performance. In terms of contemporaneous comparisons, students taking the assessment on one device can be matched to those using another device. Matching methods, like coarsened exact matching or propensity score matching are widespread and have been often used in other areas of comparability research.

Relationships Between Test Scores and External Variables

Akin to validity evidence based on relationships to external variables, finding similar associations between test scores and other criteria across devices would provide evidence of comparability. If scores produced across different computerized devices are truly measuring the same construct, it would be expected that the scores have similar correlations with external variables. Correlating test scores resulting from different devices with easily accessed criterion variables of interest (e.g., previous year achievement, grade point average [GPA], other assessment scores), can provide evidence of comparability.

Conclusion

The field of large-scale assessment is in a state of transition. Much of that transition is intentional and directly related to advances in content standards, assessment policies, and available technology. The administration of assessments on a variety of devices across different examinees is just one of the consequences of these transitions that might impact the comparability of scores. Test developers and users should understand the potential impact of the use of different devices, attempt to mitigate threats to comparability, and regularly monitor their impact. We suggest that the collection of evidence regarding comparability be routinely addressed much like other tasks that are undertaken as a regular part of the maintenance of an assessment system (e.g., year-to-year equating, new item development, or forensic analyses). Consequently, test developers and users should develop a plan to collect information on the devices used and monitor performance across devices on a regular basis. Secondary analyses should be conducted periodically—even in cases where there is little expectation of threats to score comparability.

Such investigations are particularly important given that the trends present in the current body of research may not be entirely generalizable. For example, the trends could be unique to the cohorts of students investigated, as the students of tomorrow are much more likely to have grown up using multiple devices as a routine part of their education (cf., Davis, Janiszewska, et al., 2016, p. 51). These trends could also be, in part, the result of the inclusion of grey literature. While this grey literature could be cause for concern, and potential bias, the grey literature appears to suggest differences in performance and preference just as often as the peer reviewed literature. Regardless, additional research is needed to support or refute the trends, or lack thereof, found in this work. In particular, investigations should continue to focus on the intersections between performance, fluency, and the features of assessments and devices. This type of sensitivity is also called for by Way et al. (2016), who suggest that investigations of comparability be targeted to such intersections. We also hope future studies incorporate considerations of students with disabilities, language minority students, and economically disadvantaged students. Each of these populations is likely to interact with technology differently than the general population.

To conclude, there are few straightforward questions with clear answers in large-scale assessment. Decisions made to increase reliability might negatively impact validity. Decisions made to relax standardization and increase flexibility might change the claims that can be supported or interpretations made based on the assessment results. Similarly, decisions about score comparability will come with tradeoffs. Assessment developers 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. Further, assessment scores do not exist in a vacuum—they are used to inform a number of policy decisions. Considerations of the use of assessment scores guide all of the decisions above—be they about reliability, standardization, or comparability. It is incumbent on the assessment developers and users to work together to define the uses of the assessment scores, what comparability claim supports those uses and what evidence is needed to support that claim. Finally, they must also work together to collect the needed evidence throughout the assessment cycle.

Funding

This research was supported, in part, by the Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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. doi:10.1207/S15324818AME1603_2

Chen, J., & Perie, M. (2016, April). *Comparability within computer-based assessment: Does screen size matter?* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Davis, L. L., Janiszewska, I., Schwartz, R., & Holland, L. (2016, March). *NAPLAN device effects study*. Melbourne, Australia: Pearson.

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., Morrison, K., McBride, M., & King, X. (2016, April). *Device comparability: Score range and subgroup analyses*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Davis, L. L., Orr, A., Kong, X., & Lin, C. (2015). Assessing student writing on tablets. *Educational Assessment*, *20*, 180–198. doi:10.1080/10627197.2015.1061426

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

Davis, L. L., & Strain-Seymour, E. (2013b). *Keyboard interactions for tablet assessments*. Washington, DC: Pearson Education. Retrieved 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 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), University of Kansas, Lawrence, KS.

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Part II: Online assessment in writing. In B. Sandene, N. Horkay, R. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *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 from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf

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.

Kong, X., Davis, L. L., McBride, Y., & Morrison, K. (2016, April). *Response time differences between computers and tablets*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

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

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.

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, WI: Author.

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. doi:10.1016/j.chb.2010.11.004

Sanchez, C. A., & Goolsbee, J. Z. (2010). Character size and reading to remember from small displays. *Computers & Education*, *55*(3), 1056–1062. doi:10.1016/j.compedu.2010.05.001

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. doi:10.1027/1015-5759/a000038

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.

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.

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

Table 1A. Summary of studies cited in literature review.

Study	Devices and relevant comparison(s)	Assessment(s) and subject(s)	Sample	Findings
Bridgeman et al. (2003)	Desktops (Compaq DeskPro) with 17-in. monitors set to set to (1) a resolution of 1024 × 768, (2) a resolution of 640 × 480, or (3) a simulated 15-in. monitor set to a resolution of 640 × 480	18 SAT I Verbal Items and 15 SAT I Math Items	357 college bound high school juniors	- No significant differences on math scale scores ($F(1, 309) = 2.82$, $p = .09$) - Significant differences on verbal test scores ($F(2, 309) = 3.60$, $p < .05$)
Chen and Perie (2016)	Laptops - Chromebook (14” screen) versus Mac (large, high definition screen)	ELA & Mathematics Summative Assessments	9,566 students (ELA) and 9,572 students (math) in grades 4 to 10 in 1 school district, matched on covariates	- Significant mean scale score differences in grade 4 ELA ($p < .05$), no other significant differences in grade level comparisons - 16 out of 218 ELA items had DIF, 12 out of 194 math items had DIF
Davis and Strain-Seymour (2013a)	Laptops vs. Tablets (iPads w/and w/o external keyboards)	Writing	350 Virginia students in grades 5 and 11	- More 11th grade students report using a tablet regularly in school (79%) relative to a laptop (71%), but 5th grade students report using a laptop more (84%, relative to 6%) - Few students use tablets to write essays (7% in 5th grade, 14% in 11th grade)
Davis and Strain-Seymour (2013b)	Tablets (w/ and w/o external keyboards)	Based on Strain-Seymour et al. (2013) and Davis et al. (2013)		- 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
Davis, Janiszewska, et al. (2016)	Computer (desktop or laptop; Macbook, Chromebook) vs. Tablet (with & without external keyboard; iPad, iPad Air, iPad Mini, Google Nexus, Samsung tablet)	Australia’s National Assessment Program—Literacy and Numeracy (NAPLAN)	3,602 students from 73 different schools in grades (years) 3, 5, 7, & 9	- No differences in student performance for grades 3 and 5 - Statistically significant score differences in grades 7 and 9, with students scoring higher on computers & tablets with keyboards, relative to tablets ($p < .05$) - In grade 9, there was also statistically significant score difference between tablets with keyboards and computers ($p < .05$) - Students familiar with tablets with external keyboards through classroom use perform better when tested on these devices in grades 7 and 9 - Text entry items were more difficulty on tablets across all grades - Drag and drop items were easier on tablets in grades 3 and 5, but easier on computers and tablets with keyboards in grades 7 and 9

(*Continued*)

Table 1A. (Continued).

Study	Devices and relevant comparison(s)	Assessment(s) and subject(s)	Sample	Findings
Davis et al. (2015)	Computers (Desktop and Laptops of varying make and model) vs. Tablets (full size, 9.7″, iPads running iOS 6 or higher)	59 items divided into 3 sections (reading, science, and mathematics; generally aligned to grades 7–10)	964 high school students from 5 Virginia school districts	- Mean scale scores showed no statistically significant differences between students randomly assigned to computers or tablets - Across 7 item types, only multiple select items showed statistically significant differences in performance ($p < .05$), favoring the tablet condition.
Davis, Morrison, et al. (2016)	Same as Davis et al. (2015)—Computers (Desktop and Laptops of varying make and model) vs. Tablets (full size, 9.7″, iPads running iOS 6 or higher)	59 items divided into 3 sections (reading, science, and mathematics; generally aligned to grades 7–10)	964 high school students from 5 Virginia school districts	- No significant differences between tablets and computers for math and science at any point in the score point range or for any student subgroup - Score differences in the middle to lower part of the reading score distribution (higher for tablets), driven by higher performance by male students when testing on tablets (F,1,896 = 10.184, $p < .01$)
Davis, Orr, et al. (2015)	Laptop (15″ Dell Latitude E550) vs. 10″ Tablet with onscreen keyboard vs. 10″ Tablet with external keyboard (Amazon Basic Bluetooth keyboards)	Essays prompts selected from the Pearson WriteToLearn item bank	826 students from Virginia & South Dakota in grades 5, 10 and 11	- Mean rubric scores showed no statistically significant differences - Students used computer and paper most often to write essays for course work - Students found tablet keyboards slightly more difficult to use
Davis et al., 2013	Tablets (10″ iPad (w/ and w/o external keyboard) and a 7″ Google Nexus 7)	10–15 test questions of different types from different content areas	63 students from 16 schools in Maryland, Virginia, Florida, & Texas in grade 5 and high school	- Students had more difficulty using the 7″ tablet, particularly when the area the student was trying to manipulate was small - Approximately 2/3 of students preferred the external keyboard - Most students type more text, more quickly and accurately, using the external keyboard
Eberhart, 2015	Computers (Desktop & Laptops) vs. iPad	Smarter Balanced ELA and Mathematics Assessments (3 forms per subject)	Approximately 38,000 students in grade 7	- Statistically significant main effects for device type (partial η^2 ranging from .001 to .009), as well as interactions between the effects of device type and item type for, one ELA form and all mathematics forms
		Multiple-choice (3 items per subject) and technology-enhanced items (7 items per subject)	10 students in grade 7	- Inability to select precise spots on the screen when using a tablet
Horkay et al., 2005	Laptops vs. Desktops	NAEP Eight Grade Writing Online (WOL), part of the Technology-Based Assessments study	76 students from 9 schools in grade 8 687 students in grade 8 (non-random assignment)	- No significant overall difference, but a significant difference for scores on one of two essays (F,1,72 = 4.63, $p < .05$) - No significant overall difference, but female students had higher essay scores on desktops (F,1,62 = 5.12, $p < .05$) - Essay performance is related to keyboarding proficiency, controlling for overall performance (F,1,62 = 93.40, $p < .05$).

(*Continued*)

Table 1A. (Continued).

Study	Devices and relevant comparison(s)	Assessment(s) and subject(s)	Sample	Findings
Keng et al., 2015	Computers vs. Tablets (10″ iPads with w/external keyboards)	PARCC assessments in Grades 4, 8 and 10 English Language Arts; Grades 4 and 8 Math; Geometry	286 students in grade 4 (matched to larger state sample); 267 students in grade 8; 252 students in grade 10; and 257 students in Geometry class sections (students in grades 8, 10 and geometry were randomly assigned to devices). All from one district in Massachusetts that integrated devices into classroom practice.	- Significant differences in task difficulties (when p-values or IRT difficulties displayed $\lvert z_{diff} \rvert > 1.96$ & effect sizes > 0.2) for 37% of the grade 4 math tasks, favoring computers. All other tasks had similar difficulties (3%–16% of items flagged for each assessment). - Significant difference between the correlations of the scores on the end-of-year assessment and performance-based assessment in grade 5 math (Cohen's q effect size = −0.30). No significant differences in remaining subjects and grades (q ranging from −0.08 to 0.22) - Statistically significant differences in classical reliabilities for grade 8 math ($W = 1.44$, $\Delta = 0.25$) and grade 10 ELA ($W = 1.44$, $\Delta = 0.25$). No significant differences for all other test-level reliabilities were similar ($W = 1.03$ to 1.36, $\Delta = 0.02$ to 0.21). - Statistically significant differences in correlations between scale scores and an external variable (MCAS scores) for grade 4 ELA (correlations weaker for tablets; $z_{dif} = -2.78$ & Cohen's $q = -0.21$). No significant differences for all for reaming correlations (external variable was PSAT or MCAS scores) were similar ($z_{dif} = -1.87$ to 0.84, $q = -0.21$ to −0.14). - Statistically significant differences between raw scores in grade 4 ELA, with scores lower on tablets (by 1–4 points; $\lvert z_{diff} \rvert > 1.96$). No other statistically significant differences in raw scores.
Keng et al. (2011)	Laptops (Netbooks w/10.1 inch or 11.6 inch screens) vs. Laptops (14 to 21 inch screen sizes)	Texas End-of-Course (EOC) assessments in World Geography, Geometry, and English I.	1,547 Students from 4 campuses in 2 districts	- No statistically significant mean score scale differences, holding constant the amount of information shown on screen - 7 out 127 items displayed B or C level DIF
Kong et al. (2016)	Same as Davis et al. (2015)—Computers (Desktop and Laptops of varying make and model) vs. Tablets (full size, 9.7″, iPads running iOS 6 or higher)	59 items divided into 3 sections (reading, science, and mathematics; generally aligned to grades 7–10)	964 high school students from 5 Virginia school districts	- Students took longer to respond to items on the tablet in both reading and math - No significant differences in student motivation across conditions, as indicated by the time students took to respond to item types and the tests overall
Pisacreta (2013)	Tablet vs. Tablet w/External Keboard (both	1 Typing Task, 7 Text Editing Items	10 Participants ages 11 to 13	- Typing is slower and less accurate on the Tablet vs. the Tablet w/External Keyboard - Students encountered difficulties highlighting, copying or pasting text - Students preferred the external keyboard for typing essays - Students encountered difficulty using the iPad scrolling interface

(*Continued*)

Table 1A. (Continued).

Study	Devices and relevant comparison(s)	Assessment(s) and subject(s)	Sample	Findings
Powers and Potenza (1996)	Laptops vs. Desktops	GRE Verbal, Quantitative & Essay Portions	201 undergraduate students from 9 universities	- No significant score differences on GRE Verbal and Quantitative Sections - Significant score differences on the second of two GRE Writing Prompts (higher for desktop), but only for students who responded to the first prompt on a desktop and the second on a laptop (mean *z*-score difference of 0.24)
Renaissance Learning (2013)	Computer (via a web browser) vs. iPad (via a iPad application)	STAR Enterprise assessments Reading	4,441 students in grades 1 to 11	- Scale Score correlations ranged from 0.75 to 0.92 across grades - Statistically significant scale score differences in grades 1, 5 and 8, with students taking the test on computer scoring higher ($p < .05$)
		STAR Enterprise Early Literacy	2,363 students in kindergarten to grade 2	- Scale Score correlations ranged from 0.632 to 0.868 - No statistically significant sale score differences
		STAR Enterprise Math	2,505 students in grades 1 to 11	- Scale Score correlations ranged from 778 to 0.868 - Statistically significant scale score differences in grades 1–3, 5, and 7–8 with students taking the test on computer scoring higher ($p < .05$)
Sanchez & Branaghan, 2011	Desktop (full-size screen) vs. A virtual small screen (208 x 276 pixels at 96 pip)	3 recall items & 6 application items (all multiple choice) based on two short analytical reasoning scenarios delivered as 5 e-mails	N = 34 undergraduate students	- No significant total score differences for recall items - Significant total score differences on application items, with performance in the small-display condition lower than performance in the full-size condition ($F(1, 33) = 6.81$, $p < .01$), and significant differences in response time, with the small-display condition taking longer ($F(1, 33) = 7.38$, $p < .01$)
	A virtual small screen (208 x 276 pixels at 96 pip) in portrait and landscape orientations		*N* = 33 undergraduate students	- No significant total score differences for recall items - Significant total score differences on application items, with performance landscape orientation higher than the portrait orientation ($F(1, 31) = 8.73$, $p < .01$), but no significant difference in response time - There was a significant interaction between the application total score and a measure of working memory ($F(1, 31) = 5.77$, $p < .05$), with those with higher working memory having no change in the performance, but those with lower working memory performing worse on the portrait orientation
Sanchez & Goolsbee, 2010	Desktop (19″ screen) vs. A virtual small screen device (208 x 276 pixels at 96 pip), with font size varied across texts (8 point, 8 point w/2 spaces, 12 point)	10 Item recall tasks for each of 3 expository texts on foreign countries	39 undergraduate students	- Significant total score differences, with the full-size display producing better recall than the small screen display ($F(1, 37) = 4.26$) and the 8 pt font being remembered better than the 8 pt font w/2 spaces or 12 pt font ($F(2, 74) = 8.92$, $p < .01$) - Significant interaction between display type and font size ($F(2, 74) = 3.67$, $p < .05$)

(*Continued*)

Table 1A. (Continued).

Study	Devices and relevant comparison(s)	Assessment(s) and subject(s)	Sample	Findings
Schroeders and Wilhelm (2010)	Laptops vs. Personal Digtial Assistant devices (PDAs; 3.7 inch display)	Three Reasoning Tests - Propositions, Matrices and Systems of Equations	157 high school students in Germany	- Proportion correct scores were higher when students took items on a laptop, relative to a PDA (p = .03, .43, .11 and Cohen's d = 0.19, 0.07, 0.12) - Small and uncorrelated device-specific factors for the PDA and the laptop uncovered by confirmatory factor analysis
Strain-Seymour et al., 2013	Tablet (10″ Samsung Galaxy Tab tablet w/o and w/Bluetooth External Keyboard)	Items from a Virginia Standards of Learning (SOL) field test	Twenty-four students from two Virginia school districts	- Most frequently encountered issue involved an onscreen object being smaller than or close in size to the area of a student's fingertip - Younger students and less experience typists experienced less frustration with the onscreen keyboard - The onscreen keyboard covered a portion of a typed essay - Students experienced issues with the external key board (connection, responsiveness, navigation with touchscreen)
Yu et al. (2014)	Computer (Chromebooks) vs. Tablet (iPad; ASUS & HP Android)	Performance events and multiple-choice items in three content areas— Algebra I, Biology, and English	73 students from five schools	- Student's self-rated skill with the onscreen keyboard was related to their preference for positioning of the tablet - More students prefer taking high-stakes exam on a laptop or desktop (21 students, 51%) than a tablet (17 students; 30%) - The internal keyboard and scrolling were sources of confusion for students

Table 2A. Comparisons across devices by content areas relevant to educational accountability.

Content area	No statistically significant difference in student performance	Statistically significant differences in student performance
Math	• 3rd and 5th Grade Numeracy - Davis, Janiszewska, et al. (2016) • 2nd, 6th, 10th and 11th Grade Math— Renaissance Learning (2013) • 4th to 10th Grade Math - Chen and Perie (2016) • 4th and 8th Grade Math, as well as High School Geometry (based on concorded raw scores) — Keng et al. (2015)[4] • High School Math— Davis et al. (2015); Davis, Morrison, et al. (2016) • High School Geometry—Keng et al. (2011)[3] • SAT Verbal—Bridgeman et al. (2003)[3]	• 7th and 9th Grade Numeracy— Davis, Janiszewska, et al. (2016) • 7th Grade Math— (Eberhart, 2015) • 1st to 3rd, 5th, 7th, and 8th Grade Math— Renaissance Learning (2013)
Reading (or ELA)	• Kindergarten to 2nd Grade Early Literacy— Renaissance Learning (2013) • 3rd and 5th Grade Reading - Davis, Janiszewska, et al. (2016) • 5th to 10th Grade ELA— Chen and Perie (2016) • 7th Grade ELA[2]—(Eberhart, 2015) • 3rd, 4th, 6th, 7th, 9th, to 11th Grade Reading - Renaissance Learning (2013) • 8th and 10th Grade ELA—Keng et al. (2015) • High School Reading—Davis et al. (2015) • High School English—Keng et al. (2011)[3]	• 1st, 5th, and 8th Grade Reading—Renaissance Learning (2013) • 4th Grade ELA—Chen and Perie (2016) • 4th Grade ELA—Keng et al. (2015) • 7th and 9th Grade Reading—Davis, Janiszewska, et al. (2016) • High School Reading (only at the lower end of the score distribution)— Davis, Morrison, et al. (2016) • SAT Verbal—Bridgeman et al. (2003)[3]
Writing	• 5th and High School Writing Davis, Orr, et al. (2015)[1]	• 8th Grade Writing—Horkay et al. (2005)[3] • Collegiate Writing—Powers and Potenza (1996)[3]
Science	• High School Science - Davis et al. (2015)	

Note. [1]Unlike the other comparisons in this table, Davis, Orr, et al. (2015) only compared performance on tablets to tablets with external keyboards. [2]Eberhart found nonsignificant differences on two of three forms of an ELA assessment. [3]Unlike the other studies above, these studies did not compare computers to tablets, but instead compared desktops to laptops (Horkay et al., 2005; Powers & Potenza, 1996) or differences in screen sizes, resolutions, or both (Bridgeman et al., 2003; Keng et al., 2011).[4]Keng et al. (2015) did, however, find significant differences in item difficulties for 37% of the items in grade 4 math.