Published January 2020
Research updated September 2017

CCSSO
Council of Chief State School Officers

Supplement to

Score Comparability across Computerized Assessment Delivery Devices

An update on literature produced since the June 2016 Report

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, Bureau of Indian Education, 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.

Supplement to
Score Comparability across Computerized Assessment Delivery Devices

We are grateful to our partners at the **Center for Assessment** for their support in developing this guide.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS
Jillian Balow (Wyoming), President
Carissa Moffat Miller, Executive Director

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

© 2019 by the Council of Chief State School Officers, (SUPPLEMENT TO SCORE COMPARABILTIY ACROSS COMPUTERIZED ASSESSMENT DELIVERY DEVICES), except where otherwise noted, is licensed under a Creative Commons Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/.

CONTENTS

INTRODUCTION

Any body of research evolves over time. Previous understandings become more nuanced, ideas are supported or rebuked, and, hopefully, we arrive at a clearer view of what exists before us. The research on score comparability across computerized devices is no exception. The goal of this report is to supplement *Score Comparability across Computerized Assessment Delivery Devices* (DePascale, Dadey, & Lyons, 2016) with research that has since been published or otherwise made available. This new research provides further nuance to the findings of the 2016 report, but does not change the main takeaway: Though differences in performance across devices are small, on average, and generally do not seem to follow any strong systematic trends, there are certain features of assessments and devices that have been linked to differential performance across devices, and thus may present barriers to comparability. It is worth noting up front that virtually all studies included in this supplement, old and new, compare computers—both laptops and desktops—to tablets.[1]

The 2016 report came to a number of conclusions, most of which were well summarized by Table 1 of the Appendix, which provides recommendations for states on addressing potential barriers to score comparability across computerized devices. On the next page, we provide an updated version of Table 1 from the original report with an additional column citing studies that have become available. It is clear from the distribution of the research across the recommendations in the table that recent studies on device comparability have focused more on both device fluency and specific features of the assessments (e.g., item types) and devices (e.g., screen size). However, not all of the new research falls neatly into the categories of Table 1. Much of the new research also contributes to our understanding of how device effects may differ across grade levels and subject areas. Additionally, the growing body of literature also surfaces multiple methodological advances—by providing methods that go beyond traditional studies of mean comparisons and differential item functioning (DIF).

[1] An exception is Chen & Perie (2016) that compares 14" Chromebooks to large, high-definition Macs.

Table 1

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

Recommendation	2016 Report	New
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 with 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, 2014	Davis et al., 2017b; Kong, Davis, McBride, & Morrison, 2016; Lazendic, 2017b; Steedle, McBride, Johnson, & Keng, 2016
Screen Size Establish parameters for minimum screen size. Current research suggests screens of 10” or larger reduce threats to score comparability.	Keng, Kong, & Bleil, 2011; Davis, Strain-Seymour, & Gay, 2013	Chen & Perie, 2016; Davis et al., 2017
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, 2013a	
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	Kong, Davis, McBride, & Morrison (2016)
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	
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. 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, providing students with external keyboards when responding to open-ended or composition items could support a claim of comparability.	Davis, Kong, & McBride, 2015; Davis & Strain-Seymour, 2013b; Davis, Strain-Seymour, & Gay, 2013; Pisacreta, 2013	Fitzpatrick, Tiemann, & Perie, 2017; Schwartz et al., 2017; Rabinowitz & Lazendic (2017)

DETAILS ON NEW RESEARCH

This section of our supplementary report is designed to summarize the findings associated with the new literature. To increase the interpretability of the new research, these findings are presented within the context of earlier relevant results previously provided in the June 2016 report. We do so by devoting sections to the themes outlined in the introduction. The first section addresses differences in performance by grade and subject; the second, device familiarity and fluency; the third, specific item types; and finally, the fourth section returns to the discussion of screen size from the 2016 report.

DIFFERENCES IN PERFORMANCE BY SUBJECT AND GRADE

Multiple studies published or otherwise made available since the 2016 report focus on overall student performance across multiple devices (Davis, Morrison, Kong, & McBride, 2017; Davis, Kong, McBride, & Morrison, 2017; Davis et al., 2017b;[2] Lazendic, 2017a & 2017b; Steedle, McBride, Johnson, & Keng, 2016). Generally, the results from these studies indicate that student performance, on average, is similar across computers and tablets, and that when differences in performance were present, they were generally small. These results replicate earlier findings in the June 2016 report that, in general, effects across different conditions—computers vs. tablets—are not significant (Davis, Kong, & McBride, 2015; Davis, Orr, Kong, & Lin, 2015).

Recently, several papers making these comparisons have drawn from a single large-scale study conducted using the Australian National Assessment Program—Literacy and Numeracy (NAPLAN) assessments (Davis et al., 2017b; Lazendic, 2017a & 2017b; Schwartz, et al., 2017). The NAPLAN assessment is a multi-section assessment, containing sections on reading and numeracy. The study examined student performance on the NAPLAN assessments in years (i.e., grades) 3, 5, 7, and 9 across three different device conditions—computers (laptops and desktops), tablets, and tablets with external keyboards. In both reading and numeracy in grades 3 and 5, Davis et al. (2017b) found no significant differences in mean performance across conditions. In both subjects in grades 7 and 9, the authors did find significant differences between computers and tablets, with performance favoring the computer condition. However, they also found that students who took the assessments on tablets with external keyboards performed similarly to those taking the assessments on computers—with external keyboards seemingly serving as the mitigating factor in reducing the device effect. This finding confirms earlier research that suggests external keyboards are preferable to on-screen keyboards for open-ended or composition items (Davis & Strain-Seymour, 2013b; Pisacreta, 2013).

The results of Davis, Kong, McBride, and Morrison (2017a) contrast slightly with those from the NAPLAN study. The authors examined the performance of approximately 950 high school students on a multi-section assessment of reading, science, and math. These students took the assessment either on a computer or tablet (without keyboards). The authors did not find significant differences

[2] This paper summarizes much of an earlier report by Davis, Janiszewska, Schwartz, & Holland (2016, March).

in mean performance in any grade level. Similarly, Steedle, McBride, Johnson, and Keng (2016) replicated the methods of their 2014 PARCC field test study—cited in our previous report—with the spring 2015 operational PARCC data on eight PARCC assessments (grades 5 and 7 in math, algebra 1 and 2, geometry and grades 3, 7, and 9 in English language arts) and found that, in general, student performance on tasks, correlations between the performance based assessment and end-of-year components, correlations to prior year test scores, and test reliabilities did not differ significantly between computers and tablets. The only exception was the geometry assessment, for which 11 tasks were flagged for differences in item difficulty across conditions, with performance favoring those taking the assessment on a computer. Steedle, McBride, Johnson, and Keng (2016) explained this finding as a potential indication of a lack of student familiarity or comfort in responding to geometry items on tablets. These results differ from the same study conducted using the 2014 PARCC field test data and discussed in the June 2016 report. Based on the 2014 field test data, Keng, Davis, McBride, Glaze, and Steedle (2015) found differences in the raw scores between the administrations on tablets and those on computers for the grade 4 English Language Arts assessment, the tasks for the grade 4 math assessment and geometry and, finally, reliabilities for grade 8 mathematics and grade 10 English language arts (ELA). These differences were not detected in the study of operational test data from spring 2015.

In sum, the new and previous research suggests that significant differences in overall performance (i.e., average scale scores) occur infrequently, but when such differences do occur, they occur more often in the upper grades. However, this trend of differences in the upper grades is not overly strong. In math, four studies found significant differences in student performance, and again these differences appeared more often in later grades (Davis et al., 2017b; Eberhart, 2015; Renaissance Learning, 2013; Steedle et al., 2016). Similarly, in reading or ELA, four studies found significant differences in student performance, mostly in the upper grades (Bridgeman, Lennon, & Jackenthal, 2003; Davis et al., 2017b; Keng et al., 2015; Renaissance Learning, 2013).

DEVICE FAMILIARITY AND FLUENCY

Though the studies cited above generally found few instances of significant differences in average scale scores across devices, the new literature re-emphasizes the potential importance of student-level device familiarity. The June 2016 report discussed device familiarity—or lack thereof—as a possible threat to device comparability, but at the time the number of studies formally evaluating this claim was limited. Since that time, several studies have attended to student fluency and familiarity, suggesting a shift in the literature toward more nuanced examinations. Davis et al. (2017b) found that the use of an external keyboard with the tablet mediated the differences between the computers and tablets on the NAPLAN assessments in grades 7 and 9. This strongly suggests that student facility with an onscreen keyboard is a key fluency—one that can influence student performance. Qualitative observational data confirmed that device effects can be understood by looking at particular student-device interactions (e.g., scrolling), and student familiarity with how to perform those functions fluently on the given device (Davis et al., 2017b). Using the same data, Lazendic (2017b) found that once device familiarity is controlled for, there are no significance differences in student performance across test device conditions. The results of Kong, Davis, McBride, & Morrison (2016) also provide partial support for the importance of fluency with an onscreen keyboard, as they found that students take about three to four seconds longer to answer any given item when using a tablet, relative to a computer.

ITEM TYPES

Prior research presented in the June 2016 report indicated that student performance across devices may vary by types of assessment tasks (Eberhart, 2015; Davis & Strain-Seymour, 2013a; Strain-Seymour, & Gay, 2013). In particular, technology-enhanced items were particularly susceptible to introducing differences in task performance across devices. Several additional studies have followed this line of inquiry and formally tested the interactions between specific item types and different testing devices. In general, these studies confirm previous findings—that while differences in performance at the overall scale score level tend to be small, on average, differential item functioning is present across devices.

However, the literature does not yet seem to be clearly honing in on all of the likely causes of the observed differences. Fitzpatrick, Tiemann, and Perie (2017) looked at differential item functioning (DIF) across iMacs, iPads, Chromebooks, and PCs during two years of operational statewide testing in two states that share a common item pool. In general, only a small proportion of items were flagged for DIF, and the overall effect of device on student performance appears quite small. Additionally, the researchers did not find many consistent patterns of performance across the devices by particular item features. Of the items with common identifiable features that demonstrated DIF across devices, the majority were favored in the PC testing condition. Item features that seemed to be consistently favored in the PC condition included two-column matching items and two-column multiple choice items in ELA and math, respectively, and testlets with audio passages in ELA. Item features that seemed to consistently favor iPads were ELA items with drop-down menus embedded within the text passages and ordering tasks. Though some items did demonstrate DIF in favor of iMacs and Chromebooks, there were no consistently identifiable distinguishing features on which to categorize these items.

This study was also run for students with the text-to-speech accommodation enabled. The findings for non-accommodated forms are generally replicated for this population of students; of the items that demonstrated DIF and also have consistently identifiable features, the PC seems to be the most favored device. Fitzpatrick, Tiemann, and Perie (2017) conducted follow-up cognitive laboratories with students in third, fifth, and ninth grades on all of the items flagged with DIF on the non-accommodated forms in order to better understand why particular items may favor one device over another. Based on the observational evidence, only one item type presented differences in student behavior across devices: the multiple-choice, multiple-select item type with a three-by-two layout that included images in the answer choice options. This is similar to Davis, Kong, and McBride's (2015) findings about multiple-select items, and due to the increasing popularity of these item types, more research is necessary to confirm generalizability across settings and assessments.

Similar to Fitzpatrick, Tiemann, and Perie (2017), Schwartz et al. (2017) examined differences in student performance by item type across different assessment delivery devices. Using a sample of 3,500 students across four grade levels from the NAPLAN assessment, Schwartz et al. (2017) found significant quantitative and qualitative differences across PCs, tablets, and tablets with external keyboards for both drag-and-drop and text entry item types. The device that tended to be favored for the drag-and-drop items varied by grade level, with the tablet favored for years 3 and 5 and the PC or tablet with keyboard conditions favored in the later years (7 and 9). Qualitative evidence supported that this was the most problematic item type for students on both of the tablet conditions, seemingly due to difficulties previously identified as potentially problematic in the 2016 report—e.g., small drop zones, item features close together. Similar to previous findings cited in the 2016 report, psychometric and observational data from the Schwartz et al. (2017) study also demonstrated that text entry items were generally more difficult for students taking the assessment on the tablet condition—without the external keyboard. However, in a follow up study examining NAPLAN math and spelling, Rabinowitz & Lazendic (2017) found that only about 1 percent to 3 percent of items displayed DIF between computer and tablet conditions, with no discernable pattern across item types.

To summarize both the new and old research, several item types have been shown to perform differently across devices and across settings, at least within certain assessments:

- Multiple select items (Fitzpatrick, Tiemann, & Perie, 2017; Davis, Kong, & McBride, 2015),
- Drag-and-drop items (Davis et al., 2017; Davis, Strain-Seymour, & Gay, 2013; Schwartz et al., 2017), and
- Text entry items (Davis et al., 2017b; Davis, Strain-Seymour, & Gay, 2013; Sandene et al., 2005; Powers & Potenza, 1996; Schwartz et al., 2017).

SCREEN SIZE

One of the findings of the 2016 report was related to the impact of screen size on student performance across different delivery devices. For example, Davis 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). Thus, one of the recommendations coming out of the 2016 report was for states to establish a minimum screen size of about 10" or larger in order to help minimize threats to score comparability. A new study, Chen & Perie (2016) expands upon the prior literature, not by testing how small acceptable screens may be, but by examining whether larger, high-definition screens provide any added benefit to students. This study compares student performance across two conditions, 14" Chromebooks and large, high-definition Macintosh desktops. Using propensity score matching to create two comparison groups across the conditions, Chen & Perie (2016) found little evidence that the more expensive, high-definition Mac computers provided an advantage to students as compared to standard Chromebooks—the only exception being in fourth grade ELA. This finding adds confidence to the notion that as long as states articulate a minimum acceptable screen size for test delivery, differences in performance due to screen size will likely be reduced.

METHODOLOGICAL ADVANCES

While the research above provides additional evidence supporting and adding nuance to the conclusions of the 2016 report, differences in performance across devices may or may not necessarily bear directly on comparability as the evidence necessary to support score comparability is dependent on the particular comparability claim. The 2016 report emphasizes the importance of assessment developers or users clearly articulating the comparability claim that they wish to make. The nature of the claim will then inform the *types* of evidence that should be collected to support the given claim. In the 2016 report we provided two example claims to illustrate this point: (1) If a student took the state assessment on another device, he or she would have received the same score, and (2) The student took the state assessment on the device most likely to produce the most accurate estimate of her or his true achievement.

The appendix to the 2016 report outlines a process for states to go about gathering and documenting evidence to support their intended comparability claim. The appendix provides a number of methodological suggestions for states including examining for differential item functioning (DIF analyses), comparing total test scores, analyzing internal structure, and looking at relationships between test scores and external variables. The new body of research summarized in this report draws heavily on DIF analyses, with a fair number using DIF, or DIF-like approaches, in conjunction with other approaches to examine differences in performance across devices (Chen & Perie, 2016; Fitzpatrick, Tiemann, & Perie; Lazendic, 2017a; Rabinowitz & Lazendic, 2017; Schwartz et al., 2017; Steedle et al., 2017). In addition, several new approaches, or variations on prior approaches, have emerged. These include equating independently estimated scales based on data from each device (Lazendic, 2017a); examining distributions of performance across devices, both overall and for subgroups, (Davis, Morrison, Kong, and McBride, 2017); and applying linear mixed-effect models to capture variation across and interactions between device, student familiarity, students, and items (Lazendic, 2017b).

Lazendic (2017a) scaled item responses from each device separately, equated these separate scales individually to the paper version, and then compared the resulting linked item parameters across assessments. If item calibrations are the same or similar across the different equating procedures, this provides robust evidence of score comparability across different delivery devices. Of course, an assumption of this method that the sample of linking items is adequately representative of all the tested items.

Davis, Morrison, Kong, and McBride (2017) examine differences in score distributions across devices, as opposed to the much more common comparisons of means. This method is based on the Matched Samples Comparability Analysis (MSCA) developed by Way, Davis, and Fitzpatrick (2006), which was developed to examine score differences between paper and computer-delivered assessments. Because students could be randomly assigned to different device conditions, Davis, Morrison, Kong, and McBride (2017) did not need to use matching methods to create equivalent samples. This method is similar to Lazendic (2017a) in that equating was used to test differences in item calibration, but instead of equating different device conditions to a third paper-based condition, raw score to raw score equating was used to directly compare the computer and tablet conditions. If the results show that the expected tablet raw score for students taking the assessment on a computer is the same as that raw score for the students taking the assessment on a tablet, this indicates that performance does not generally differ across the devices in question.

Lazendic (2017b) contributes another methodological tool to examine differences across devices by applying a linear mixed-effect model to account for different sources of variation in scores, such as device type, device familiarity, and item type. This type of modeling allows for direct testing of the significance of the assessment delivery device in explaining variation in student scores. A non-significant main effect for the device variable provides evidence of score comparability across different devices. The three studies discussed here contribute important methodologies that states should consider adding to their toolboxes as they are designing their plans for collecting evidence related to the comparability of scores across delivery devices.

CONCLUSION AND RECOMMENDATIONS

Collectively, the current body of research suggests that students generally perform similarly across computerized devices. However, the research did find there can be some exceptions to this rule—exceptions that can often be traced back to student familiarity with a device or to specific types of items. As the body of research has begun to focus more closely on the specific aspects of devices or assessments that cause differences in performance, the methods have followed suit, becoming more sensitive and focused on detecting differences beyond the average overall scale score.

Because the implementation of statewide assessments across computerized devices is still relatively new, and therefore the peer-reviewed academic literature is limited, it is strongly recommended that any state supporting the delivery of their assessment system across multiple computerized devices continue to

1. attempt to mitigate the potential impact of the use of different devices,
2. document evidence of score comparability, and
3. regularly monitor for any detectable effects of the delivery device on the state's claim of comparability.

States can minimize potential barriers to score comparability by engaging in thoughtful design and development of the assessment and by establishing clear guidelines during assessment implementation. During the assessment development phase, states can design and test items to ensure they render and function similarly across all approved assessment delivery devices and conduct cognitive labs to provide evidence that students engage in similar cognitive processes when interacting with the items across the devices. During both the design implementation phases, states can use the research provided in the June 2016 report and this supplement to make informed decisions regarding devices, specifications, tools, and items for minimizing threats to score comparability. Additionally, states can establish clear protocols that define key aspects of the testing program, such as a list of approved devices, clear administration procedures and training materials, and plans for continued quality assurance.

In addition to minimizing potential barriers to score comparability, states should be incorporating data collection and analysis plans for gathering evidence of score comparability whenever multiple devices are used to deliver an assessment. This evidence can come from documentation related to test design and development such as reports resulting from the types of cognitive labs described in the previous paragraph. Evidence can also come in the form of the programmatic documentation such as feedback from trainings or practice materials or documentation of device-related incident reports during an assessment administration. The most compelling evidence of score comparability, however, is likely to come from post-administration data analyses—including comparisons of overall scale scores, examinations of item response data for differential item function, or any of the new methods described in the methodological advances section of this supplement.

Lastly, states should plan for ongoing monitoring of potential barriers to score comparability for both commonly used and newly introduced test delivery devices. Although the bulk of evidence may be collected during the introduction of a new device, states should plan to continue to collect evidence for continued assurance of score comparability. Just as states continuously monitor for construct-irrelevant variance associated with the use of assessment accommodations, for the foreseeable future states will need to monitor assessment results for continued evidence of device comparability—to help guard against ongoing and newly emerging threats to the validity of the interpretations derived from the assessment scores.

REFERENCES

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.

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: 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., Kong, X., McBride, Y., & Morrison, K.M. (2017a). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education*, *30*(1), 16-26, DOI: 10.1080/08957347.2016.1243538.

Davis, L.L., Morrison, K., Kong, X., & McBride, Y. (2017b). Disaggregated effects of device on score comparability. *Educational Measurement: Issues and Practice, 36*(3), 35-45.

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

Davis, L.L., Schwartz, R., Janiszweka, I., Holland, L., Businovski, B., & Lazendic, G. (2017, April). *Evaluation of device effects in the NAPLAN online assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

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

Davis, L.L., & Strain-Seymour, E. (2013b). *Keyboard interactions for tablet assessments.* Location unknown: Pearson. Retrieved June 21, 2018, from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/CRPAF34829_CA-Flyer_Keyboard_final_web.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*. Location unknown: Pearson.

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Washington, DC: CCSSO. Retrieved June 21, 2018, from http://www.ccsso.org/Documents/CCSSO%20TILSA%20Score%20Comparability%20Across%20Devices.pdf.

Eberhart, T. (2015). *A comparison of multiple-choice and technology-enhanced item types administered on computer versus iPad* [KU Doctoral dissertation]. Retrieved June 21, 2018, from https://kuscholarworks.ku.edu/bitstream/handle/1808/21674/Eberhart_ku_0099D_14325_DATA_1.pdf?sequence=1&isAllowed=y.

Fitzpatrick, J., Tiemann, G., & Perie, M. (2017, February). *Item comparability across different electronic assessment devices*. Paper presented at the winter meeting of the Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards, New Orleans, LA.

Keng, L., Davis, L., McBride, Y., Glaze, R., & Steedle, J. (2015). *Spring 2014 digital devices comparability research study*. Report for the 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.

Lazendic, G. (2017a, April). *The impact of test devices on equating of online and paper tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Lazendic, G. (2017b, April). *Application of LMEM to evaluate device effects in NAPLAN*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Lorié, W. (2015, March). *Reconceptualizing score comparability in the era of devices.* IGNITE presentation at the annual conference of the Association of Test Publishers.

Pisacreta, D. (2013, June). *Comparison of a test delivered using an iPad versus a laptop computer: Usability study results.* Paper presented at the Council of Chief State School Officers' National Conference on Student Assessment, 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.

Rabinowitz, S. & Lazendic, G. (2017, April). *Analysis of device effects in the context of multistage adaptive NAPLAN tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

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

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.

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 from http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf

Schwartz, R., Davis, L.L., Janiszweka, I., Holland, L., Businovski, B., Choen, A., Traecy, K., & Lazendic, G. (2017, April). *Interaction of device effects and item type in the NAPLAN online assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Steedle, J., McBride, M., & Johnson, M. & Keng, L. (2016). *Spring 2015 digital devices comparability research study* [White paper]. Location unknown: Pearson.

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]. Location unknown: Pearson. Retrieved June 21, 2018, from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/002_Testing-on-Tablets-PartI.pdf.

Way, W.D., Davis, L.L., & Fitzpatrick, S.J. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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 *Technology and Testing: Improving Educational and Psychological Measurement* (F. Drasgow, Ed.). NYC, NY: Routledge.

CCSSO

Council of Chief State School Officers

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