

Device Comparability of Tablets and Computers for Assessment Purposes

National Council on Measurement in Education
Chicago, IL

Laurie Laughlin Davis, Ph.D.

Xiaojing Kong, Ph.D.

Yuanyuan McBride, Ph.D.

April, 2015

Abstract

The definition of what it means to take a test online continues to evolve with the inclusion of a broader range of item types and a wide array of devices used by students to access test content. To assure the validity and reliability of test scores for all students, device comparability research should be conducted to evaluate the impact of testing device on student test performance. The current study looked at the comparability of test scores across tablets and computers for high school students in three commonly assessed content areas and for a variety of different item types. Results indicate no statistically significant differences across device type for any content area or item type. Student survey results suggest that students may have a preference for taking tests on devices with which they have more experience, but that even limited exposure to tablets in this study increased positive responses for testing on tablets.

Keywords: tablets, mode comparability, device comparability, score comparability

Device Comparability of Tablets and Computers for Assessment Purposes

Use of digital devices for instruction and assessment continues to increase as Bring Your Own Device (BYOD), 1:1 technology programs, and flipped learning change the way students are interacting with academic content, with their teachers and peers, and demonstrating their mastery of academic knowledge and skills (Hamdon, McKnight, McKnight, & Arfstrom, 2013; Johnson, 2012; McCrea, 2011; Ballagas, Rohs, Sheridan, & Borchers, 2004). According to the Speak Up 2013 national survey results of K-12 students (Speak Up, 2013) the most common academic uses for devices include looking up information on the internet (63% of students report this use), accessing online textbooks (43% of students report this use), communicating with peers and teachers (42% of students report this use), and taking an online test (40% of students report this use).

Online testing, as the fourth most common academic use of devices, merits specific attention as the definition of what it means to take an online test continues to evolve and diversify. Online testing has moved beyond the traditional multiple choice item as the primary vehicle for evaluating student competencies. Instead, this generation of tests includes a variety of different “technology-enhanced” item types (TEIs for short) that seek to provide more authentic measurement of student skills with reduced guessing and more use of constructed response. Additionally the devices used by students to access the online test content vary considerably across classrooms and school districts. Desktop computers with large monitors and external keyboards and pointing devices (aka mice) may exist side-by-side in school districts with more portable laptop computers which have smaller keyboards, monitors, and a variety of different pointing device options. Laptop computers, in turn, may exist side-by-side

in school districts with even more portable tablet devices which have yet smaller screens and keyboards (onscreen or external) and use fingers as the pointing device.

The increasing diversity of technology in schools raises the question of whether all devices are created equal relative to how students can use them to demonstrate their knowledge and skills when taking an online test. Professional testing standards (APA, 1986; AERA, APA, NCME, 2014, Standards 9.7 & 9.9) require evidence that neither the mode of delivery nor the device used to access test content should influence the interpretations of students' scores and assessment outcomes. Thus it is important to evaluate the comparability of scores across devices. Way, Davis, Keng, and Strain-Seymour (in press) suggest that "form factor" of the device is an important consideration when evaluating comparability. The "form factor" describes the way in which the student uses the device to access and manipulate digital content—the more similar the form factor of two devices the more comparable scores which result from testing on those two devices can be expected to be. Desktop computers and laptop computers both have similar form factors and research has shown that student performance across these devices is relatively comparable (Keng, Kong, and Bleil, 2011; Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, and Oranje, 2005; Bridgeman, Lennon, & Jackenthal, 2001; Powers & Potenza, 1996). However, there are significant differences in the form factors between computers and touch-screen tablets that may influence student experience in using the devices, and, in turn, their resulting test scores. Specifically the screen size of tablets (typically 7-inch to 10-inch) is smaller than that of computers which may limit how much information students can see on screen at one time. Additionally, the method of input for student responses on tablets relies on using the finger to select and move information on screen and either the onscreen keyboard or an external keyboard to enter text.

While there have been a number of qualitatively conducted cognitive labs and usability studies looking at student interactions with tablets (Yu, Lorié, & Sewall, 2014; Piscreta, 2013; Lopez & Wolf, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013; Davis, Strain-Seymour, & Gay, 2013) the quantitative research evidence in this area is more limited. Davis, Orr, Kong, and Lin (in press) found no effects of device (tablet vs. computer) for student writing of short essays for either 5th grade or high school students. Similarly Olsen (2014) concluded that there were no effects of device (tablet vs. computer) for a set of multiple choice items in reading, writing, and mathematics across a K-12 grade span set of tests. As both the PARCC and Smarter-Balanced assessment consortia are allowing students to use tablets in taking their tests across a wide variety of item types the need for further research in this area is evident (PARCC, 2013; SBAC, 2013). As such, the current study looked at the comparability of student test scores across tablets and computers for three commonly assessed content areas (reading, mathematics, and science) and for a variety of different item types (multiple choice and TEIs).

Method

Participants

Data were collected in spring 2014 from a sample of 964 high school students from five different school districts in Virginia—Frederick, Henrico, Isle of Wight, Prince George, and Stafford. Student participants were required to have completed or be currently enrolled in coursework in Algebra I, Biology, and English II by the time of the study. All students who participated in the study had prior experience with taking tests online as part of the Virginia Standards of Learning (SOL) assessment program. Schools were offered incentives based on the proportion of eligible students who participated in the study. Schools that provided 50% or

more of their eligible students for study participation were provided with one iPad and cover. Schools that provided 75% or more of their eligible students for study participation were provided with two iPads and covers. Recruitment of participating schools was conducted with support of the Virginia Department of Education. Rather than obtaining a permission slip for each student in the study, schools were asked to notify parents and legal guardians of the study and to follow an opt-out procedure for student permissions. Some schools additionally followed their own local policies with regard to obtaining permission slips for students to participate in the study.

A total of 964 students submitted responses to the test. Table 1 shows the demographic characteristics students participating in the study broken out by study condition.

Table 1: Demographic characteristics of participants

	Tablet	Computer
Male	258 (53.2%)	248 (51.8%)
Female	224 (46.2%)	227 (47.4%)
Missing	3 (0.6%)	4 (0.8%)
White	310 (63.9%)	305 (63.7%)
African-American	105 (21.6%)	87 (18.2%)
Hispanic	28 (5.8%)	33 (6.9%)
Other	36 (7.4%)	46 (9.6%)
Missing	6 (1.2%)	8 (1.7%)
TOTAL	485	479

Apparatus/Materials

Hardware

Researchers arrived at each school location and set-up the study equipment prior to beginning the study sessions. Each school participated in both the computer and tablet conditions. Specific set-up requirements differed across school locations, but, in general,

students in the computer condition participated in computer labs or classrooms using school provided computers. Computers for this study included a mix of desktop and laptop models with the only specifications being that they meet the requirements for running the testing software. For some schools, students had been issued their own computers (1:1 device assignment with laptops) whereas for other schools students worked on computers in computer labs.

For the majority of schools, tablets were rented and provided for the study so that data collection could be conducted efficiently across a one to two day period with up to 100 students testing at one time. One school (Smithfield High School in Isle of White) had issued tablets to each of their students (1:1 device assignment with tablets) and students used their own tablets for the study. In all cases, tablets were required to be full size (9.7") iPads running iOS 6 or higher. For rented tablets, study facilitators installed and configured software in advance of the study session. For student provided tablets, this process was conducted under the direction of study facilitators prior to beginning the test session. For this study tablets were used without peripherals (e.g. styluses, stands, or external keyboards). Students used the onscreen keyboard to provide brief typewritten responses to fill-in-the-blank questions.

Figure 1 provides photos that shows the configuration of a testing room in one participating school where both computer and tablet were set-up within the same room.



Figure 1: Example classroom set-up for data collection

Software

Students accessed the test content through the online testing software application. The software used in this study was accessible by computer via any web browser without special software installation. To access the software from the tablets an application had to be downloaded so a small amount of additional set-up was needed. Once the software is launched (from either the computer or the tablet) it locks out all other applications to allow for secure test administration. The tools offered within the software were limited for this study and only included a four-function calculator (which should have been sufficient to answer all math and science items on the test). Additionally, students did have access to flag and review functionality which allowed them to flag an item for review and navigate back to that item from later points in the test. Due to timing of the study and the version of software available, no test-taking tools (e.g. answer choice eliminator, highlighter, etc.) were available. Students were, however, given scratch paper and pencils to support any interim work or calculations they wanted to record.

Measures

Each student in the study responded to a set of 59 items divided into 3 sections (reading, science, and mathematics) and a short set of survey questions about their experiences. Only items which could be computer or algorithmically scored (i.e. short fill-in items) were included. Study content was selected from a variety of sources and was not intended to reflect or explicitly align with a specific set of content standards or test blueprint. Reading passages were selected to reflect a range of genres and lengths. In mathematics, item selection targeted computation, geometry, and pre-algebra strands. The majority of content was selected to reflect skills which were somewhat below grade level (grade range 7-10) for participating high school students so that students could reasonably be expected to have learned the material. Table 1 shows the item allocation across content area and item type for the study.

Table 1. Test Blueprint for the Study

Content Area	Item Type							TOTAL
	Multiple Choice	Hot Spot	Drag and Drop	Fill In the Blank	Multiple Select	Inline Choice	Graph Point	
Reading	6	4	8	0	2	0	0	20
Mathematics	8	0	1	4	4	0	1	18
Science	12	2	2	2	0	3	0	21
TOTAL	26	6	11	6	6	3	1	59
% of Test	44%	10%	19%	10%	10%	5%	2%	100%

The test was sequenced so that students completed the reading section first, followed by the science section, and finally the math section. In between sections students were prompted with a “motivation” screen (see figure 2) designed to provide positive reinforcement for their effort in the study (Bracey, 1996; Brown and Walberg, 1993).



Figure 2: Example motivation screen

Procedures

Students were randomly assigned to condition either in advance of the study (based on classroom assignment) or at the time of the study (students were alternately assigned either to tablet or computer conditions). At the beginning of each study session, a facilitator introduced themselves, briefly discussed the purpose of the study, provided directions to the students about what to do, and answered any questions. The version of the software used for this study differed from the version students had used previously for online testing, however, study facilitators reviewed functionality such as navigation and tools with the students prior to the beginning of the test session. Students were then given 80 minutes to read and respond to the test items. Following completion of all three subject area sections, students were asked to complete a 10-question survey about their home and school use of different devices as well as their experience in the study itself.

Results

Although random assignment to condition was employed throughout the data collection process, state reading test scores were obtained and used as a check on the effectiveness of the random assignment. Table 2 shows a comparison of scale scores on the state high school

reading assessment for students in the tablet and computer conditions. Mean scores are very similar and differences were not statistically significant.

Table 2: Mean score by content area

Study Condition	N*	State Reading Score
Tablet	384	449.73 (40.12)
Computer	367	451.36 (32.62)

*Note that reading scores were not available for all students who participated in the study.

Motivation Filtering

Motivation filtering (Wise, Kingsbury, Thomason, & Kong, 2004; Sundre & Wise, 2003) was conducted based on student response times within each section of the test. Motivation filtering was applied to remove student participants from the data when it was clear that their responses were provided too quickly to support accurate score interpretations.

Descriptive information about student response time for each section is presented in Table 3. Response time information was examined separately for students who were in the computer and tablet conditions but was very similar for both conditions. Based on review of the response time information, students who spent less than four minutes responding to the reading section (n=8), less than two minutes responding to the science section (n=11), and less than two minutes responding to the math section (n=47) were excluded from further analysis.

Table 3: Response time by content area

Content Area	Response Time (in Minutes)			
	Mean (SD)	Min	Max	N
<i>Tablet</i>				
Reading	18.90 (6.64)	3.15	68.63	485
Science	16.93 (6.35)	1.22	37.52	484
Mathematics	9.80 (4.16)	0.97	32.00	478
<i>Computer</i>				
Reading	17.93 (7.35)	1.52	64.33	479
Science	15.50 (6.75)	0.12	36.33	477
Mathematics	9.04 (4.29)	0.10	31.78	472

Content Area and Item Type Analysis

Multiple choice items were scored as 0/1(wrong/right) as was the single graph point item on the test form. All other item types were scored to allow for partial credit where applicable. Points were assigned for full credit (worth 2 points), partial credit (worth 1 point), or no credit (worth 0 points) as indicated in figure 3 where n is the number of correct components within an item for Group 1 items and the number of components within an item for Group 2 items, and Sum reflects the sum of +1 for all correctly selected components and -1 for all incorrectly selected components. Additional details on the scoring rules for each item can be found in Appendix A.

Group	Item Type	Scoring Rule Applied
1	<ul style="list-style-type: none"> • Multiple select • Hot spot • Drag and drop (single bay or single dragger) 	Fully correct --> 2 Sum = n-1 --> 1 Else --> 0
2	<ul style="list-style-type: none"> • Fill in the Blank • Inline choice • Drag and drop (multiple bays) 	Fully correct --> 2 Correct picks $\geq n/2$ --> 1 Else --> 0

Figure 3: Scoring rules for assigning partial credit

The primary inferential analysis was an independent samples t-test conducted for the total score within each of the three content areas (across all item types) and within each item type (across all content areas). Student responses to each item were scored and aggregated within content area to produce a total score for reading, science, and mathematics. Additionally, responses were aggregated by item type across all content areas to produce a total score for each item type. Table 4 provides the descriptive statistics for total score in each content area. Means across study condition are very similar and t-tests showed no statistically significant differences (alpha level 0.01) between student performance on tablets and computers.

Table 4: Mean score by content area

Content Area	Study Condition	N-count	Mean Score	Standard Deviation	Mean % of Total Points	Statistical Test
Reading	Tablet	483	19.5	5.04	57.36	t=1.35; p=.1767
	Computer	473	19.05	5.25	56.03	
Science	Tablet	481	12.61	4.74	42.04	t=0.27; p=.7900
	Computer	472	12.53	4.46	41.78	
Mathematics	Tablet	466	8.96	5.69	31.99	t=-0.34; p=.7369
	Computer	451	9.08	5.56	32.44	

Table 5 provides the descriptive statistics for total score by item type. Means across study condition are very similar and t-tests showed no statistically significant differences (alpha level 0.01) between student performance on tablets and computers.

Table 5: Mean score by item type

Item Type	Study Condition	N-count	Mean Score	Standard Deviation	Mean % of Total Points	Statistical Test
Multiple Choice	Tablet	465	15.24	4.93	58.60	t=0.24; p=0.8127
	Computer	448	15.16	4.73	58.31	
Drag and Drop	Tablet	465	11.72	3.72	53.26	t=0.54; p=0.5866
	Computer	448	11.58	3.74	52.65	
Hot Spot	Tablet	465	4.04	2.15	33.66	t=-0.72; p=0.4745
	Computer	448	4.14	2.06	34.49	
Fill in the Blank	Tablet	465	3.18	1.79	26.51	t=0.83; p=0.4072
	Computer	448	3.08	1.86	25.67	
Multiple Select	Tablet	465	3.81	2.51	31.72	t=2.04; p=.0412
	Computer	448	3.47	2.47	28.91	
Inline Choice	Tablet	465	3.05	1.19	50.75	t=-0.25; p=0.8050
	Computer	448	3.06	1.21	51.08	
Graph Point	Tablet	465	0.37	0.48	18.60	t=-1.13; p=.2596
	Computer	448	0.41	0.49	20.42	

Item Analysis

Individual item means were divided by the total number of points available for each item to create a p-value on a scale of 0.0-1.0 for each item. Item p-values for the tablet

condition were then compared to item p-values for the computer condition using an independent samples t-test. Figure 4 shows the p-value differences for items within each content area. Across all 59 items only 2 items (both within the reading content area) showed statistically significant differences at the $p < 0.01$ level favoring tablet—items 7 ($t=2.77$; $p=.0056$) and 18 ($t=2.60$; $p=.0093$). Both items were of multiple choice item type.

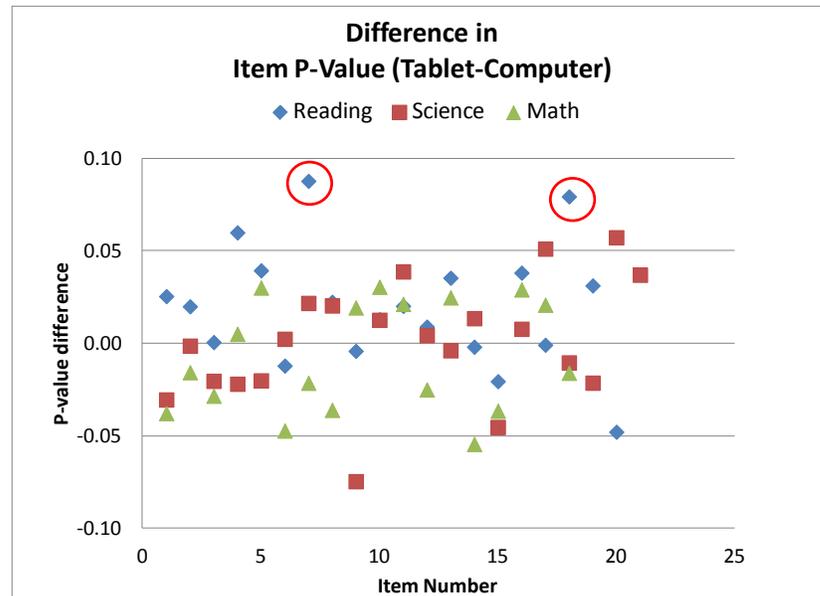


Figure 4: Differences in item p-value by content area

Survey Analysis

Figures 5 through 11 show selected responses to the survey questions. A total of 962 high school students responded to the survey. However, student response rates to individual questions varied somewhat, therefore percentages are reported in relationship to the number of students responding to each question and do not include missing responses. For some survey questions, students were asked to choose all responses that applied so percentages will total to more than 100%.

As seen in figures 5 and 6 more students in this study report using tablets and smart phones for personal use than for school work (44% vs. 28% for tablets and 84% vs. 31% for

smart phones) though rates of usage for desktop and laptop computers are more similar between personal use and school work (27% vs. 36% for desktop computers and 61% vs. 66% for laptop computers). Figure 7 shows information about students' previous use of devices specifically for the purpose of taking tests. Not surprisingly, most students reported previous experience taking tests on paper (95%) and on computer (both desktop at 85% and laptop at 75%), but few had experience taking a test on a tablet (24%) or smart phone (5%).

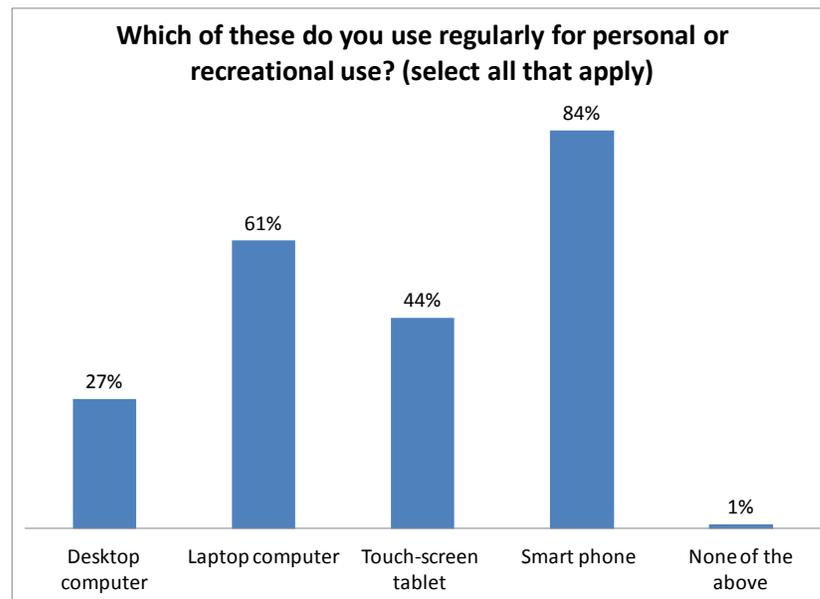


Figure 5: Device use for personal/recreational (n=961)

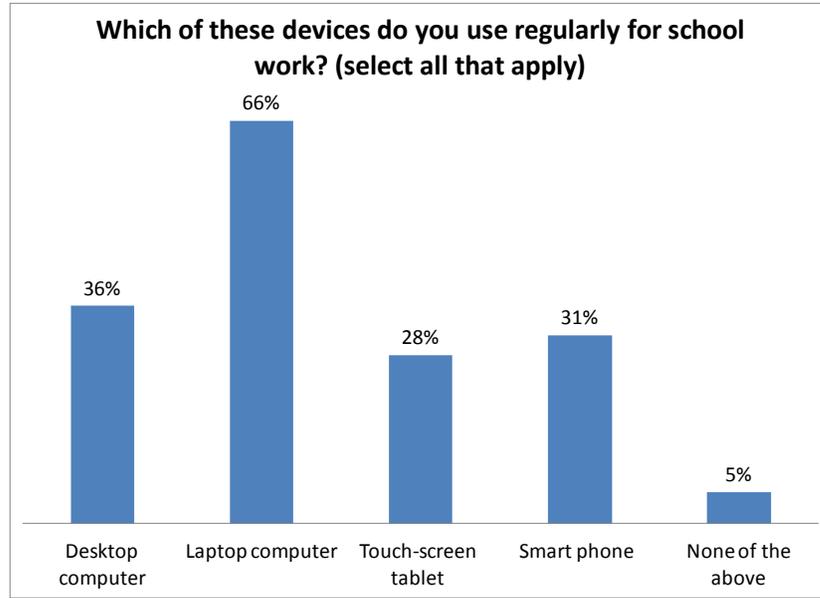


Figure 6: Device use for school work (n=960)

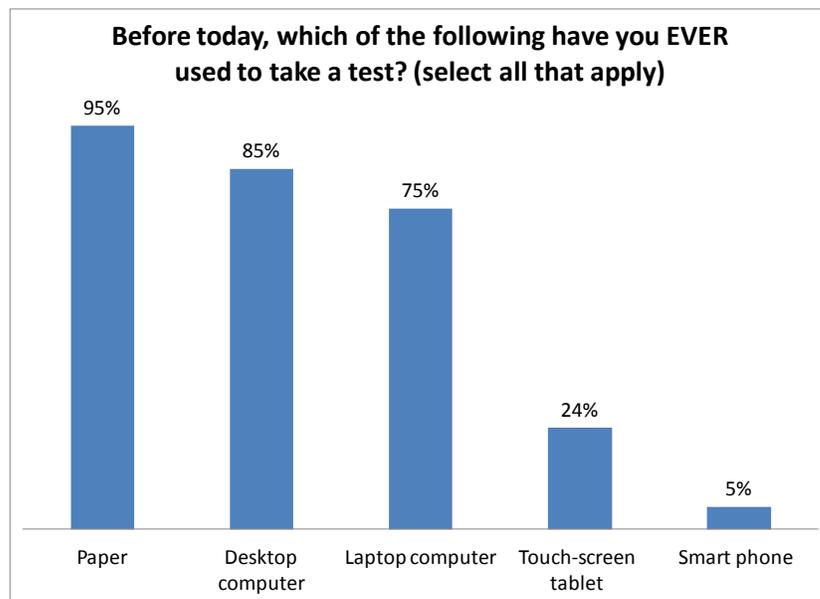


Figure 7: Previous experience with devices for taking tests (n=960)

Figure 8 shows the response of students broken out by study condition when asked what mode or device type they preferred for taking a test. Interesting the most frequently endorsed options were paper only (selected by 30% of students in the computer condition and 22% of students in the tablet condition) or paper and computer (selected by 27% of students in the computer condition and 18% of students in the tablet condition). Across both study

conditions, few students indicated that they would prefer to take a test on a touch-screen device (such as a tablet or smart phone). However, those who participated in the tablet condition of the study tended to select those options at a higher rate than those who participated in the computer condition (e.g. 14% vs. 7% for touch-screen only; 13% vs. 9% for computer and touch-screen). As such it appears that the experience with taking a test on tablet during the study may have positively impacted student perceptions toward the device.

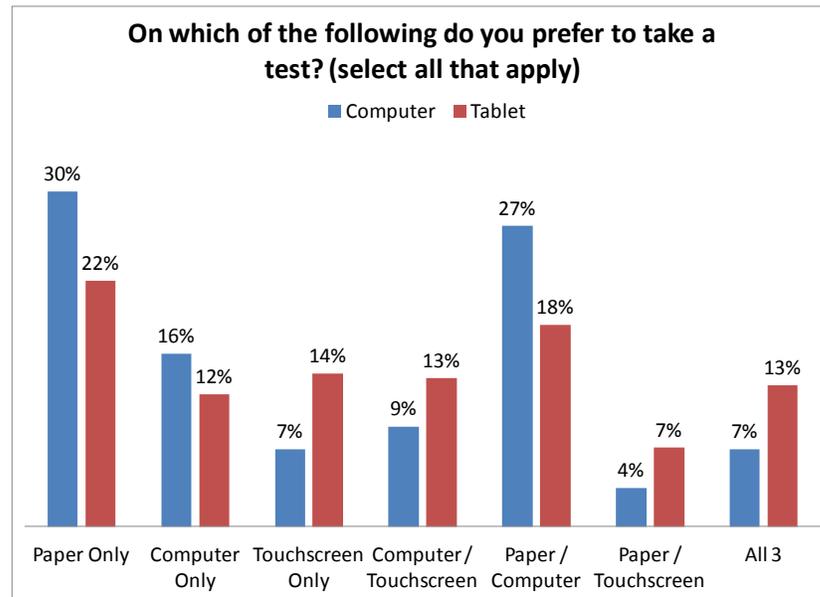


Figure 8: Devices preference for taking tests (n=959)

Figures 9 through 11 show student perceptions of the difficulty of questions in each of the three content area sections (reading, science, and math) broken out by study condition. For all three content areas, students in the tablet condition were more likely to respond that they thought the questions were either “very easy” or “easy” than students in the computer condition. This was most pronounced for reading where 58% of students in the tablet selected “very easy” or “easy” compared with 44% of students in the computer condition. Differences were somewhat smaller for science (49% for tablet vs. 40% for computer) and math (41% for tablet vs. 37% for computer). This suggests that the experience with taking a test on tablet

during the study may have also positively impacted student perceptions toward the test questions themselves.

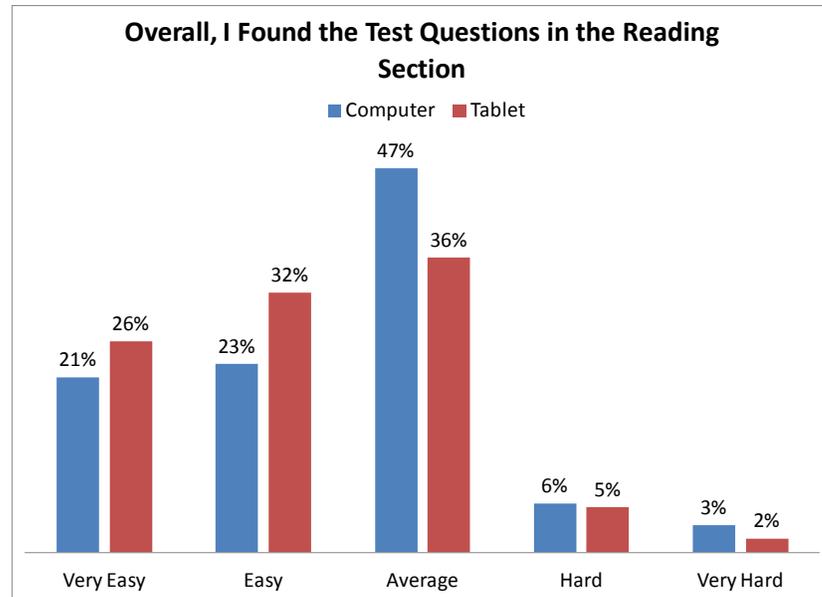


Figure 9: Question Difficulty by Study Condition for Reading (n=958)

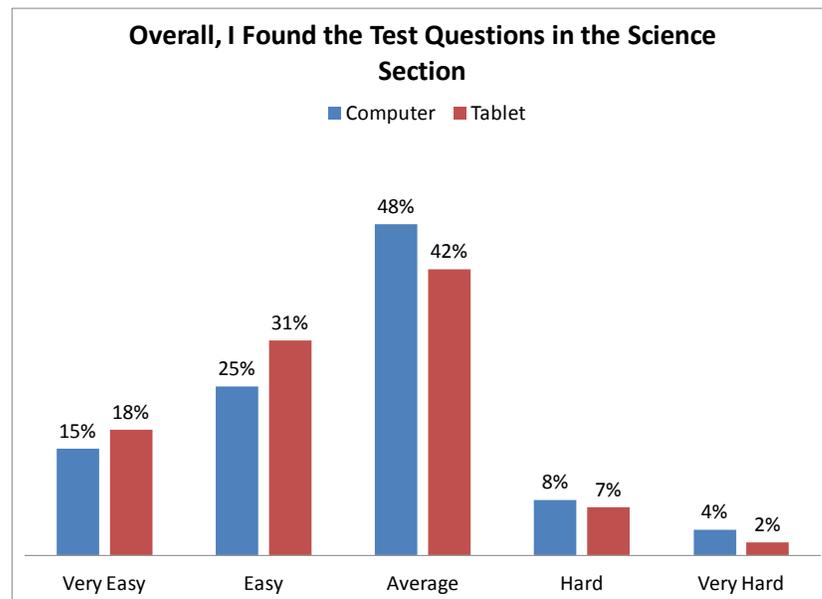


Figure 10: Question Difficulty by Study Condition for Science (n=957)

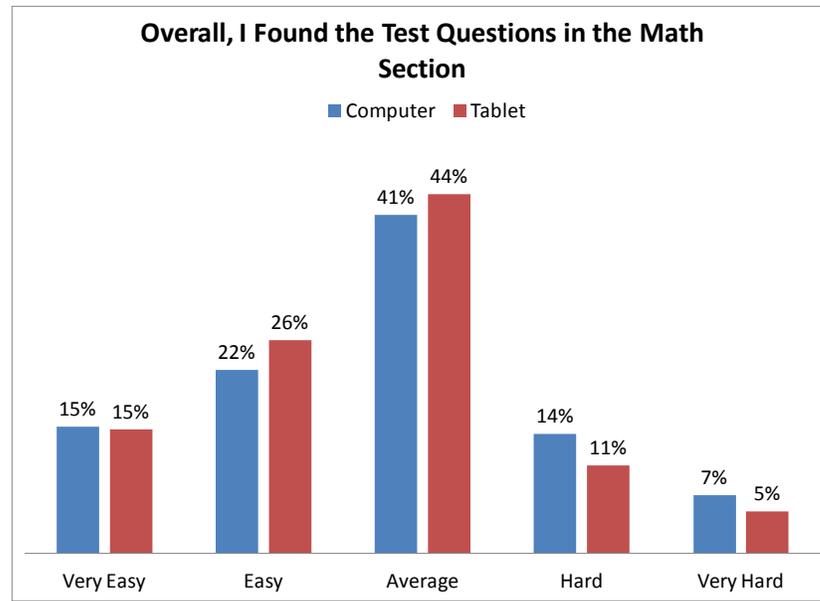


Figure 11: Question Difficulty by Study Condition for Math (n=956)

Additional Qualitative Analysis

In addition to the formal survey efforts conducted across all study participants, one participating school division (Prince George; n=322 students—half of whom would have tested in the tablet condition) captured comments from their students after the study and summarized these as “likes” and “dislikes” about testing on tablets. These comments were used by the study researchers to create word clouds (as seen in Figures 12 and 13) with larger words indicating more frequent mention within the student comments and smaller words indicating less frequent mention within the student comments.

Table 6. Summary of Themes from Tablet “Like” and “Dislike” Word Clouds

Tablet Likes	Tablet Dislikes
Easy/Easier	Uncomfortable
Reading/Read	Small
Navigate	Glare
Look, touch, feels	Stand
New	Fingers, Neck
Convenient	Sitting
Zoom	Hurt
Practical	Typing
Modern	Math

Discussion

Overall the results from this study are consistent with the findings of previous studies in this area (Davis, Orr, Kong, and Lin, in press; Olsen, 2014). There were no observable performance differences in student test scores across device for any of the three content areas or across any item type evaluated (though admittedly the number of observations for any given item type varied considerably with some item types having a very small number of items included in the study). In many ways this is reassuring as use of tablets for large scale testing has moved forward with the PARCC and Smarter-Balanced assessments as well as for many individual state next generation assessment programs. The findings of this study should not, however, be interpreted to suggest that tablets should only be used as a means of expanding the set of available devices on testing day. Student familiarity with tablets in an academic context is crucial and tablets are best used as part of a technology rich learning environment throughout the school year.

The pattern of high school student preferences observed in this study relative to preference of device for testing is consistent with previous findings (Davis, Orr, Kong, & Lin, in press; Strain-Seymour, Craft, Davis, & Elbom, 2013; Davis, Strain-Seymour, & Gay, 2013). Older students tend to be more aware of the uses and consequences associated with test scores

than younger students (where preference for tablets tends to be greater) such that they prefer to use devices with which they have more experience. However, this study did demonstrate that students who had experienced testing on the tablet as part of this research held somewhat more positive views of using tablets for testing than those who did not have that opportunity. This suggests that with additional experience and practice using tablets for test-taking (such as through tutorials or practice test sessions) that student comfort levels with tablets may further improve.

Despite the lack of statistically significant differences in student performance for reading, there is some evidence in the survey and qualitative results to suggest that reading may be an area where students prefer to work with tablets. The scrolling interface used to present the reading passages provides for a very natural gesture with the finger on the touch-screen device whereas the use of the mouse as an intermediary device to scroll the passage on the computer may be somewhat more cumbersome. Additional tablets and eReaders are becoming more and more common such that students may find interacting with text on a touch-screen device a very familiar activity. However, it should be noted that the reading passages presented in the current study were relatively short. Different performance and preference outcomes might be observed if longer reading selections were studied. Future research might look specifically at student interactions with reading on tablets to understand more about these factors.

Lastly, this study used a single size (10-inch form factor) of tablet for data collection. This decision was deliberate and reflects current practice relative to allowable device sizes on large scale assessment programs. However, many school districts have smaller tablets (7-8-inch, 5-inch, or even 3-4-inch smart phones) that students use in daily classroom activities.

Future research might consider the comparability of student performance across a range of tablet and touch-screen sizes. This might allow districts to test their students with the same devices they use in classroom activities, and might encourage districts who currently do not use tablets at all to consider purchasing smaller, more cost effective tablets if they can be used for both instructional and assessment purposes.

References

- American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC.
- American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments (APA) (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.
- Ballagas, R., Rohs, M., Sheridan, J. & Borchers, J. (2004). BYOD: Bring Your Own Device. *UbiComp 2004 Workshop on Ubiquitous Display Environments*, September, Nottingham, UK.
- Bennett, R.E.(2003). *Online Assessment and the comparability of score meaning* (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.
- Bracey, G.W. (1996). *Altering the motivation in testing*. *Phi Delta Kappan*, 251-252.
- Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS-RR-01-23). 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(3), 191-205.
- Brown, S.M & Walberg, H.J. (1993). *Motivational Effects on Test Scores of Elementary Students*. *Journal of Educational Research*, 86 (3), 133-136.
- Davis, L.L., Orr, A., Kong, X., & Lin, C. (in press) Assessing student writing on tablets. *Educational Assessment*.
- 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*. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-Part-II_formatted.pdf
- Hamdon, N., McKnight, P., McKnight, K., & Arfstrom, K.M. (2013). A review of flipped learning. Retrieved from http://www.flippedlearning.org/cms/lib07/VA01923112/Centricity/Domain/41/LitReview_FlippedLearning.pdf
- Johnson, D. (2012). Power up! On board with BYOD. *Educational Leadership*, 70 (2), pages 84-85. Retrieved from: <http://www.ascd.org/publications/educational-leadership/oct12/vol70/num02/On-Board-with-BYOD.aspx>

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, Louisiana

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

Lopez, A, & Wolf, M.K. (2013, December). *A Study on the Use of Tablet Computers to Assess English Learners' Language Proficiency*. Paper presented at the annual meeting of the California Educational Research Association, Anaheim, CA.

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

McCrea, B. (2011). Evolving 1:1: THE Journal. Retrieved from:
<http://thejournal.com/articles/2011/05/11/evolving-1-to-1.aspx>

Olsen, J.B. (2014, April). *Score comparability for web and iPad delivered adaptive tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Partnership for the assessment of Readiness for College and Careers (2013, February). *Technology Guidelines for PARCC assessments version 2.1 – February 2013 Update*. Retrieved from
http://www.parcconline.org/sites/parcc/files/PARCCTechnologyGuidelines2dot1_Feb2013Update.pdf

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 (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 Report No. RR-96-15) Princeton, NJ: Educational Testing Service.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., and Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from:
<http://nces.ed.gov/nationsreportcard/pdf/studies/2005457.pdf>

Smarter Balanced Assessment Consortium (SBAC 2013, February). *The Smarter Balanced technology strategy framework and system requirements specifications*. Retrieved from
http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Technology-Strategy-Framework-Executive-Summary_2-6-13.pdf

Speak Up (2013). *The new digital learning playbook: Understanding the spectrum of students' activities and aspirations*. Retrieved from <http://www.tomorrow.org/speakup/pdfs/SU13StudentsReport.pdf>.

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*. Retrieved from <http://researchnetwork.pearson.com/wp-content/uploads/Testing-on-Tablets-PartI.pdf>.

Sundre, D.L., & Wise, S.L. (April, 2003). 'Motivation Filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics tests*. San Antonio, Texas: Harcourt.

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

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

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (in press). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In *Technology in testing: Measurement issues*, ed. F. Drasgow. Vol 2 of the NCME book series.

Wise, S.L., Kingsbury, G.G., Thomason, J., & Kong, X. (April, 2004). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Winter, P. (2010). *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers: Washington, DC. Retrieved from: http://www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf

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 A: Scoring Rules by Item Type

- MC items: 1 for correct answer and 0 for wrong answer.
- Inline choice items:
 - If there is one inline question in the item, it could be treated as a MC item;
 - If there are multiple (n) inline questions in the item:
 - Each answer is either scored as 1 or 0.
 - ✓ If the sum is less than $n/2$, code the score as incorrect (0);
 - ✓ If the sum is equal to or larger than $n/2$ but less than n , code the score as partially correct (1);
 - ✓ If the sum is equal to n , code the score as correct (2);
 - ✓ The possible scores for this item are 0, 1, and 2.
- Fill-in-the-blank items:
 - If there is one fill-in-the-blank question in the item, it is scored as either correct (1) or incorrect (0).
 - If there are multiple (n) fill-in-the-blank questions in the item:
 - Each answer is either scored as 1 or 0.
 - ✓ If the sum is less than $n/2$, code the score as incorrect (0);
 - ✓ If the sum is equal to or larger than $n/2$ but less than n , code the score as partially correct (1);
 - ✓ If the sum is equal to n , code the score as correct (2);
 - ✓ The possible scores for this item are 0, 1, and 2.
- Multiple-select and Hot-spot items:
 - Score each correct pick as 1 and each incorrect pick as -1 and add these up.
 - If the sum is less than $n-1$, code the score as incorrect (0, n is equal to the number of correct answers, same below);
 - If the sum is equal to $n-1$, code the score as partially correct (1);
 - If the sum is equal to n , code the score as correct (2);
 - The possible scores for this item are 0, 1, and 2.
- Graph-point items:
 - 1 for correct answer and 0 for wrong answer.
- Drag-and-drop items:
 - Some drag-and-drop items in this study can be treated as multiple-select items
 - One bay with several draggers;
 - Multiple bays with one dragger (drag and drop the dragger to the appropriate bay(s));
 - The rest of the drag-and-drop items include: (1) Multiple bays with several (more than the number of bays) draggers: drag and drop one dragger for each bay and the bays are identical. (2) “matching game”: There are multiple bays (each of which is unique) and multiple draggers in the item and the number of draggers could be equal to or higher than the number of bays. The testers are required to match each dragger to one of the bays. There are two situations for matching:

each bay can only be matched to one of the dragger or each bay can be matched to multiple draggers.

- Each dragger that is put in the correct bay is scored as 1, otherwise 0;
- If the sum is less than $n/2$ (n is equal to the number of draggers, same below), code the score as incorrect (0);
- If the sum is equal to or larger than $n/2$ but less than n , code the score as partially correct (1);
- If the sum is equal to n , code the score as correct (2).
- The possible scores for this item are 0, 1, and 2.