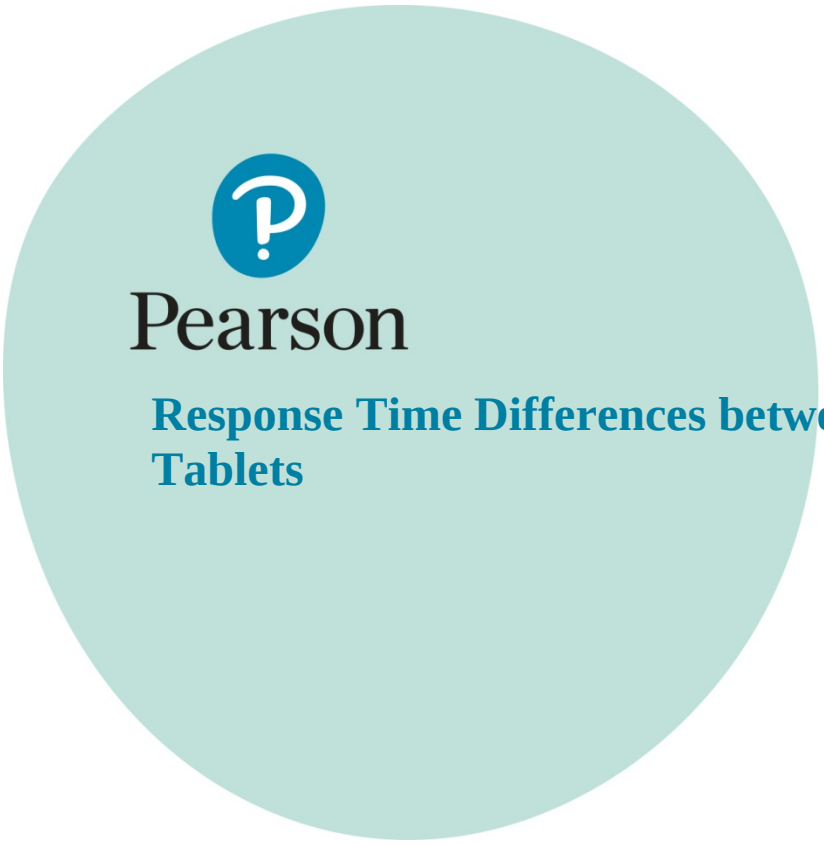

Response Time Differences between Computers and Tablets

National Council on Measurement in Education
Washington, D.C.

Xiaojing Kong, Ph.D., Pearson

Laurie Laughlin Davis, Ph.D., Pearson

Yuanyuan McBride, Ph.D., Pearson

Kristin Morrison, M.S., Georgia Institute of Technology

April, 2016

Abstract

Item response time data were used in investigating the differences in student test-taking behavior between two device conditions: computer and tablet. Analyses were conducted to address the questions of whether or not the device condition had a differential impact on student motivation (with response time effort being used as an indicator) as well as on the time that student spent on the test (reading, mathematics, and science) or a given item type (e.g., drag-and-drop and fill in blank). Further analyses were conducted to examine if the potential impact of device conditions varied by gender and ethnicity groups. Overall, there were no significant differences in student motivation across conditions. Students tended to spend more time when taking the mathematics and science tests, as well as certain types of items, on the tablet. Follow-up research on the item time thresholds is discussed.

Keywords: tablets, response time, motivation, response time effort

Response Time Differences between Computers and Tablets

The use of tablets for digital assessment delivery has gained increasingly more attention over the past several years (PARCC, 2013; SBAC, 2013; MDE, 2015). This follows on the heels of a shift in technology use in classrooms as schools have embraced the use of tablets in support of blended and flipped learning (e.g. Hamdon, McKnight, McKnight, & Arfstrom, 2013) to allow students access to a growing body of digital content. A number of research studies have evaluated the use of tablets for assessment, including usability studies and cognitive labs (Pisacreta, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013; Davis, Strain-Seymour, & Gay, 2013; Yu, Lorie, & Sewall, 2014) as well as comparability studies (Olsen, 2014; Keng, Davis, McBride, & Glaze, 2015; Davis, Orr, Kong, & Lin, 2015; Davis, Kong, McBride, & Morrison, in press). Although some have revealed differences in the ways in which students interact with touch-screen tablets when compared with computers (e.g. the use of the finger as a pointer, the onscreen keyboard, etc.), a consistent finding has been that different testing devices (computer vs. tablet) do not show significant impact on student test performance (i.e., non-statistically significant and/or non-practically significant effects at the total score level).

While lack of statistically and/or practically significant impacts to student test performance is encouraging in terms of score comparability, it is possible that the use of different devices may impact student interactions with test items in other ways. For example, direct interaction with the finger on a touch-screen device may be faster and more efficient for student interactions than using a mouse. Similarly, students may be more engaged and, therefore more motivated, when using touch-screen devices because of their ability to interact directly with objects on the screen.

In comparing device inputs, a contrast can be made between the immediacy of input when using a finger to directly manipulate objects on a touch-screen and the accuracy of input when a mouse or touch-pad is used. Moving objects directly onscreen more closely resembles interactions in the physical world and may seem more natural to students than using a mouse to achieve intended outcomes. Additionally, speed and intuitiveness are both benefits of direct input, especially for novices. However, use of a mouse improves accuracy beyond what is possible with the human finger alone (Hinckley & Wigdor, 2011). Touch inputs are typically faster than mouse inputs for working with large objects onscreen (e.g., greater than 3.2 mm), but are less accurate than mouse inputs for working with objects smaller than 10.5 mm (Vogel, Baudisch, & Shift, 2007; Hall et al., 1988; Sears & Shneiderman, 1991; Meyer, Cohen & Nilsen, 1994; Forlines et al., 2007).

In terms of student motivation and engagement, Davis, Strain-Seymour, and Gay (2013) found that 5th grade and high school students reported they would like to take a test on a tablet. Students in this study found the experience to be more interactive and hands-on, thought that the graphics were sharper and clearer than if viewed on a computer, and described the experience as being “more fun” than testing on either computer or with paper-and-pencil. However, there may have been some novelty effects represented in these findings, which could be expected to dissipate over time as tablet use in the classroom and assessment becomes more commonplace. In fact, Project Tomorrow (2015) reports that middle school students identified a laptop as the best device for taking an online test regardless of whether they regularly used laptops or tablets in the classroom. This finding ran counter to the findings for several other digital tasks (e.g. taking notes, reading an online textbook, etc.), where preference for device tracked closely with device familiarity.

Item response times can be used to evaluate both differences in student motivation and engagement, as well as speed and efficiency of interaction. Item response times are a valuable piece of information that can be conveniently collected during a computer based test and have not yet been studied in the context of evaluating comparability between computers and tablets. This data can be very useful in investigating examinee test-taking behavior. As Wise and Kong (2005) pointed out, there are several advantages of using response time data: (a) collection of response time is unobtrusive and non-reactive; (b) unlike examinee self-report, response time information represents a direct observation of examinee behavior; (c) since response time information is available for each item, it permits investigations of dynamic changes in examinee behavior during a testing session.

The term *response time*, as used in this paper, is generally defined as the differences in time (seconds) between item presentation and when a response is made by the examinee. Previous research has explored the use of response time in many areas, including speededness issues (Schnipke & Scrams, 1997), obtaining more accurate proficiency and recovering IRT item parameter estimates (Thissen, 1983; Wise & DeMars, 2006), identifying appropriate time limits (Bhola, Plake, & Roos, 1993; Gershon, Bergstrom, & Lunz, 1993), detecting item pre-knowledge and cheating (Impara, Kingsbury, Maynes, & Fitzgerald, 2005; Meijer & Sotarido, 2006; Qian, Staniewska, Reckase, & Woo, 2016; van der Linden & Gao, 2008), and selecting items in computerized adaptive testing (CAT; van der Linden, 2008). For research focusing on low-stakes testing, item response times have been particularly useful in the detection of solution behavior and rapid guessing behavior (Wise & DeMars, 2006; Wise & Kong, 2005; Wise & Ma, 2012).

Examinees who provide good effort to a test item will exhibit solution behavior. In contrast, examinees who do not try to do well on a test item will exhibit rapid-guessing

behavior. Rapid-guessing behavior occurs when an examinee responds to an item so quickly that he or she could not have adequate time to read and fully consider the item (Schnipke & Scrams, 1997; Wise & Kong, 2005). This phenomenon was first examined in large-scale high stakes testing, and known as speededness. In low-stakes testing, however, the lack of student motivation seems a more plausible explanation for the occurrence of rapid-guessing behavior.

In the current research, response time data were used in investigating the differences in student test-taking behavior between two different testing conditions (computer vs. tablet) in a low-stakes testing environment. More specifically, the following research questions were addressed:

- Was there a difference between devices in student engagement and motivation?
- Was there a difference between devices in the time that students spent on the test?
- Was there an interaction between device and student subgroup (gender or ethnicity) in terms of either student engagement/motivation or time students spent on the test?
- Was there a difference between devices in terms of either student engagement/motivation or time students spent on different item types?

Method

Participants

Response data were collected in spring 2014 from a sample of 964 high school students from five different school districts in Virginia. Each school participated in both the computer and tablet conditions. 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. Table 1 shows the demographic characteristics of the students participating in the study broken out by study

condition. Additional information about the school and student participants can be found in Davis, Kong, McBride, & Morrison (in press).

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. The test was sequenced so that students completed the reading section first, followed by the science section, and finally the math section. Student responses to each item were scored 0, 1, or 2 based on a previously determined set of partial credit scoring rules (see Davis, Kong, McBride, & Morrison, in press). Table 2 shows the item and point allocation across content area and item type for the study.

Hardware and Software

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 the majority of schools, tablets (9.7" iPads running iOS 6 or higher) 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 tested at one time. 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 set-up was needed. Additional information about the study hardware and software can be found in Davis, Kong, McBride, & Morrison (in press).

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 previously used 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. Response time for each item was captured as the difference between the initial display of each item and the navigation to the next item. If students went back to review an item after initially responding, this time was not included as part of an item's response time.

Data Analysis

Two sets of analyses were conducted to explore differences between computer and tablet conditions in terms of differential impact on motivation and the time spent.

Set 1. Motivation/Effort Analyses

To analyze student engagement and motivation, response time effort (RTE) was used. This was initially proposed in Wise and Kong's study (2005). Procedurally, for item *i*, there is a threshold, T_i, that represents the response time boundary between rapid-guessing behavior and solution behavior. Given an examinee *j*'s response time, RT_{ij}, to item *i*, a dichotomous index of item solution behavior, SB_{ij}, is computed as

$$SB_{ij} = \begin{cases} 1 & if \;\; RT_{ij} \geq T_i, \\ 0 & otherwise. \end{cases}$$

The index of overall response time effort for examinee *j* to the test (or a given item type) is given by

$$RTE_j = \frac{\sum SB_{ij}}{k},$$

where *k* is the number of items in the test/item type. RTE scores range from zero to one, and represent the proportion of test items for which the examinees exhibited solution behavior. RTE values near one indicate strong examinee effort to the test, and the farther a value falls below one, the less effort the examinee expended.

During the computation of RTE values, a critical step is to set an item time threshold. According to Wise and Ma (2012), a 10% normative threshold (NT10) is recommended, as this method markedly outperformed a common three-second threshold used in previous research. For example, if it takes students an average of 40 seconds to respond to a particular item, the response time threshold for this item would be 4 seconds (that is, 10% of 40 seconds). A response time equal to or greater than 4 seconds would be classified as solution behavior, while a response time of less than 4 seconds would be classified as rapid-guessing behavior. A maximum threshold value of 10 seconds was set, because it would be problematic to credibly characterize a response taking longer than 10 seconds as a rapid guess (Wise and Ma, 2012). Item time threshold was determined separately for the computer and tablet conditions.

Once the RTE values were obtained, a two-factor analysis of variance (ANOVA), using device as one factor and gender/ethnicity as the second factor, was conducted to look for potential effects of device on motivation and/or interactions between device condition and student subgroup. Additionally, an independent samples *t*-test was conducted to look at differences in motivation across device conditions by item type.

Set 2. Time Analyses

Instead of RTE, raw response times were used in this set of analyses. Similar to the first set of analyses, a two-factor ANOVA was performed using device and gender/ethnicity as the two independent variables. RTE values were used to filter out students with low motivation from the dataset prior to analysis of raw response times. This ensured that differences in response time between devices would reflect differences in speed or efficiency of device usage rather than being confounded with differential effects of student motivation by device. Additionally, an independent samples *t*-test was conducted to look at differences in response time across device conditions by item type.

Results

Motivation/Effort Analyses

Out of 59 items, 40 (68%) of them had the same thresholds for both device conditions. For the remaining 19 items, the absolute difference was 1 second. After the thresholds were applied, each student and item combination was classified as either solution behavior or rapid guessing behavior. Across the 59 items, the percentages of solution behavior ranged from 86% to 100% for the tablet condition and from 81% to 100% for the computer condition; the percentage of rapid-guessing behavior ranged from 0% to 14% for the tablet condition and from 0% to 19% for the computer condition. Item score accuracies were used as validity checks to provide evidence that the recommended threshold-identification method was effective in separating rapid-guessing behavior from solution behavior. The score accuracies were computed for separately for students who were classified as having solution behavior for an item and those who were classified as demonstrating rapid guessing behavior on an item. For the multiple-choice items, the score accuracy for the rapid guessing group was at the chance level overall, while for the technology-enhanced items, the score accuracy for the rapid guessing group was much lower than that for the solution behavior group.

ANOVA (Device x Gender)

Table 3 shows the mean RTE values for students in each device condition by gender for reading, science, and math, respectively. For reading, there were no statistically significant differences for either main effect ($F(1, 946)=0.50$, $p>.01$, for device condition; $F(1,946)=5.59$, p>.01 for gender) or the interaction between device condition and gender, $F(1,946)=1.39$, $p>01$). For science, there were no statistically significant differences for either the interaction between device condition and gender ($F(1, 943)=5.07$, $p>.01$) or the main effect ($F(1,943)=0.86$, $p>.01$ for device condition. The main effect for gender, $F(1, 943)=18.06$, $p<.01$, was statistically significant. Similarly, for mathematics, the interaction between device condition and gender was not significant ($F(1, 933) = 0.28$, $p>.01$). The main effect for device condition ($F(1, 933)= 0.12$, $p>.01$) was not significant either. However, the main effect for gender, $F(1, 933)=12.25$, $p<.01$, was statistically significant. As the purpose of this research is to understand the effect of device on subgroups rather than to evaluate differences within the subgroups themselves, the significant differences for gender will not be further interpreted.

ANOVA (Device x Ethnicity)

Table 4 shows the mean RTE values for students in each device condition by ethnicity for reading, science, and math, respectively. For all three content areas, there were no statistically significant differences for either the interaction between device condition and ethnicity (reading: $F(2,944)=1.15$, $p>.01$; science: $F(2, 941)=0.50$, $p>.01$; math: $F(2, 931)=0.88$, $p>.01$), or any of the main effects including the main effect of devices (reading: $F(1,944)=2.22$, $p>.01$; science: $F(1, 941)=0.26$, $p>.01$; math: $F(1, 931)=0.10$, $p>.01$) and the main effect of ethnicity (reading: $F(2,944)=0.48$, $p>.01$; science: $F(2, 941)=2.83$, $p>.01$; math: $F(2,931)=0.60$, $p>.01$).

t-test for Item Types

Tables 5 shows the mean RTE values for students in each device condition and item type. There were no significant differences found in student motivation and effort between device conditions.

Time Analyses

The RTE values computed from the first set of analyses were used as a motivation filter to remove student records that showed a high percentage of rapid guesses. Only students with a RTE value of 0.75 or higher were included in the time analyses. Across device conditions, students took an average of 18-19 minutes to complete the 20 items in the reading section, 16-17 minutes to complete the 21 items in the science section, and 9-10 minutes to complete the 18 items in the math section.

ANOVA (Device x Gender)

Table 6 shows the mean response time values (in seconds) for students in each device condition by gender for reading, science, and math, respectively. For reading, there were no statistically significant differences for either main effect ($F(1, 938)=4.87$ $p>.01$, for device condition; $F(1,938)=1.85$, $p>.01$ for gender) or the interaction between device condition and gender, $F(1,938)=0.88$, $p>01$. For science, there were no statistically significant differences for either the interaction between device condition and gender ($F(1, 893)=5.61$, $p>.01$) or the main effect for gender ($F(1,893)=1.181$, $p>.01$). The main effect for device, $F(1, 893)=16.39$, $p<.01$, was statistically significant, with students testing on tablets spending more time on the test students testing on computer. For mathematics, the interaction between device condition and gender was not significant ($F(1, 855) = 0.19$, $p>.01$). However, both main effects were statistically significant (device: $F(1, 855)= 11.59$, $p<.01$; gender: $F(1, 855) = 24.30$, $p<.01$). Students in the tablet condition had longer response times than the computer condition. As the purpose of this research is to understand the effect of device on subgroups rather than to

evaluate differences within the subgroups themselves, the significant differences for gender will not be further interpreted.

ANOVA (Device x Ethnicity)

Table 4 shows the mean response time values (in seconds) for students in each device condition by ethnicity for reading, science, and math, respectively. The ethnicity subgroup analyses showed that for reading, there were no statistically significant differences for either the interaction ($F(2, 936)=0.22$, $p>.01$) or the main effect for testing devices ($F(1,936)=3.37$, $p>.01$). The main effect for ethnicity, $F(2, 936)=30.29$, $p<.01$, was statistically significant.. For science, there were no statistically significant differences for either the interaction ($F(2, 891)=0.31$, $p>.01$) or the main effect for ethnicity ($F(2,891)=3.40$, $p>.01$).Statistical significant difference was found for the main effect of device condition ($F(1,891)= 9.47$, $p>.01$), with students spending more time when taking the science test on the tablet. For mathematics, there were no statistically significant differences for either the interaction effect ($F(2,853)=0.32$, $p>.01$) or for the main effect of ethnicity ($F(2,853) =1.27$, $p>.01$). The main effect for devices was statistically significant ($F(1,853) = 7.37$, $p<.01$). As the purpose of this research is to understand the effect of device on subgroups rather than to evaluate differences within the subgroups themselves, the significant differences for ethnicity will not be further interpreted.

t-test for Item Types

Table 8 shows the mean response time values (in seconds) for each item type. Note that the number of items of each type varied, with most of the items (26) reflecting multiple choice and smaller numbers of items (1 to 11) for each of the technology enhanced item types. Statistically significant differences in response time between tablet and computer conditions were found for all item types except drag and drop. For all other item types, students in the

tablet condition took longer to respond than students in the computer condition (averaging 3.48 seconds longer per item for multiple choice items, 4.12 seconds longer per item for hot spot items, 3.59 seconds longer per item for fill in the blank items, 3.53 seconds longer per item for multiple select items, 7.68 seconds longer for inline choice items, and 3.44 seconds longer for the graph point item). The effect sizes were small (Cohen, 1988), ranging from 0.19 to 0.29.

Discussion

Overall, this study found no evidence for differential levels of student motivation or effort related to device. Although the study allowed for different RTE thresholds to be determined for each item by device condition, the thresholds were identical for most items and very similar for the remaining items. Additionally, differences in test-level RTE values were not significant for either the main effect of device or the interaction of device with student subgroups for any of the three content domains. Lastly, there were no significant differences in RTE found for any item type (across domains). This may not ultimately be that surprising a finding for high school students. Although Davis, Strain-Seymour, and Gay (2013) found some evidence that students would prefer taking a test on tablet, this preference was most frequently expressed by younger students. Older students tended to have a more developed understanding of test consequences and therefore expressed more concerns about how use of a tablet would impact their test-taking experience. Additionally, the idea that preference for taking a test on tablet will lead to increased motivation is likely moderated by a number of factors, including the strength of the preference, the difficulty of the test content, and the stakes of the testing situation.

Conversely, there were differences observed between device conditions for raw response time. Students in the tablet condition took significantly longer to respond to items

than students in the computer condition for both science and math. Differences in response times were not dramatic, however. They averaged about 1 minute 40 seconds longer for the science section and about 52 seconds longer for the math section. It is worth noting that these differences in response times did not lead to any differences in student performance on the test itself (see Davis, Morrison, McBride, & Kong, 2016; Davis, Kong, McBride, & Morrison, in press). However, these differences do suggest that the immediacy and directness of using the finger as the input device did not result in expected gains in speed or efficiency in responding. Instead, it appears that the reduced precision resulting from using the finger as the input device rather than a mouse may have created a small degree of challenge for working with the onscreen objects within the science and math tests.

A review of response time results by item type can further inform this discussion. Students in the tablet condition took significantly more time to respond to items than students in the computer condition for all item types except for drag and drop. The difference was in the neighborhood of 3-4 seconds per item for most item types. For inline choice items, the difference was a bit larger (nearly 8 seconds longer for tablet than for computer), but the number of items of this type evaluated in this study (N=3) was quite small. Caution should be used to avoid over-generalizing this result. The lack of a significant difference for drag and drop items is somewhat surprising, since this item type can be particularly sensitive to precision issues relative to the size of draggable objects and accuracy issues relative to placing those objects in drop zones. However, the directness provided by using the finger to drag and drop objects onscreen more closely matches the directness of moving objects in the real world and, therefore, may offer sufficient advantages in terms of the naturalness of this interaction that the challenges with accuracy and precision may be partially overcome.

One of the limitations of the study falls with the determination of thresholds for RTE. While the NT 10 rule seemed to work reasonably well for the majority of the items, there were two multiple choice items for which the score accuracy rates for the rapid-guessing group were very close to or slightly higher than the score accuracy rates for the solution behavior group. This brings into question whether the thresholds for these items are, in fact, appropriately separating rapid guessing from solution behavior. Other methods for determining thresholds should therefore be considered. Additionally, student raw response times and response time efforts decreased across the three sections of the test with the longest response times and greatest RTE values observed for reading (average of 18-19 minutes with average RTE values of 0.99), followed by those for science (average of 16-17 minutes with average RTE values of 0.95), and lastly by those for math (average of 9-10 minutes with average RTE values of 0.93-0.94). Clearly motivation and fatigue played some role in the study results though perhaps not differentially by device. It is unknown, however, how results would have differed if the math or science sections were administered first rather than the reading section.

References

Bhola, D. S., Plake, B.S., & Roos, L. L. (1993, October). *Setting an optimum time limit for a computer-administered test.* Paper presented at the annual meeting of the Midwestern Education Research Association, Chicago, IL.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Davis, L, L. Kong, X., McBride, Y., & Morrison, K. (In press). Device comparability of tablets and computers for assessment purposes. *Applied Measurement in Education.*

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

Davis, L.L., Strain-Seymour, E., & 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

Forlines, C., Wigdor, D., Shen, C., & Balakrishnan, R. (2007, May). *Direct-touch vs. mouse input for tabletop displays.* Paper presented at CHI, San Jose, CA.

Gershon, R. C., Bergstrom, B. A., & Lunz, M. (1993, April). *Computer adaptive testing: Exploring examinee response time.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Hall, A.D., Cunningham, J.B., Roache, R.P., & Cox, J.W. (1988). Factors affecting performance using touch-entry systems: Tactual recognition fields and system accuracy. *Journal of Applied Psychology*, 4, 711-720.

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

Hinckley, K. & Wigdor, D. (2011). Input Technologies and Techniques. In Andrew Sears and Julie A. Jacko (eds), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 161-176). CRC Press.

Impara, J.C., Kingsbury, G., Maynes, D. & Fitzgerald, C. (2005) Detecting cheating in computer adaptive tests using data forensics. Paper presented at the annual meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.

Keng, L., Davis, L.L, McBride, Y. & Glaze, R. (2015). *PARCC spring 2014 digital devices comparability research study.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Meijer, R. R. & Sotaridona, L. S. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing* (LSAC Computerized Testing Report No. 03-03). Newtown, PA: Law School Admission Council.

Meyer, S., Cohen, O., & Nilsen, E. (1994, April). Device comparisons for goal-directed drawing tasks. In *Conference companion on Human factors in computing systems* (pp. 251-252). ACM.

Minnesota Department of Education (2015, May). *Minnesota Tablet Usability Study Report.* Retrieved from http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowlege in compute-based licensure examinations. *Educational Measurement: Issues and Practice, 00(0),* 1-10.

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.

Project Tomorrow (2015). Digital Learning 24/7: Understanding Technology-Enhanced Learning in the Lives of Today's Students. http://www.tomorrow.org/speakup/SU14DigitalLearning24-7_StudentReport.html

Sears, A., & Shneiderman, B. (1991). High precision touchscreens: design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies*, *34*(4), 593-613.

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

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34,* 213-232.

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.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5-20.

van der Linden, W.J., & Guo, F. (2008). Bayesian Procedures for Identifying Aberrant Response-Time Patterns in Adaptive Testing. *Psychometrika*, *73*(*3*), 365-384.

Vogel, D., & Baudisch, P. (2007, April). Shift: a technique for operating pen-based interfaces using touch. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 657-666). ACM.

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

Way, W.D., Davis, L.L., Keng, L., & Strain-Seymour, E. (2015). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow, (Ed.), *Technology and testing: Improving educational and psychological measurement* (NCME Applications of Educational Measurement and Assessment, Vol. 2, pp. 260-284). New York: Routledge.

Wise, S.L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19-38.

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. Applied Measurement in Education, 18, 163-183.

Wise, S.L., & Ma, L. (2013). Setting response time thresholds for a CAT item pool: The normative threshold method. Paper presented at the 2012 annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

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.

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%)
Black	105 (21.6%)	87 (18.2%)
Other	64 (13.2%)	79 (16.5%)
Missing	6 (1.2%)	8 (1.7%)
TOTAL	485	479

Table 2. Test Blueprint for the Study

Item Type								
Content Area	**Multiple Choice**	**Hot Spot**	**Drag and Drop**	**Fill In the Blank**	**Multiple Select**	**Inline Choice**	**Graph Point**	**Total # of Items (Points)**
Reading	6	4	8	0	2	0	0	20 (34)
Mathematics	8	0	1	4	4	0	1	18 (23)
Science	12	2	2	2	0	3	0	21 (28)
TOTAL	26	6	11	6	6	3	1	59 (85)
% of Test	**44%**	**10%**	**19%**	**10%**	**10%**	**5%**	**2%**	**100%**

Table3. Mean RTE by Device Condition and Gender

Reading					Science					Mathematics				
Device Condition	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**
Tablet	Female	0.99	0.06	222	Tablet	Female	0.96	0.10	221	Tablet	Female	0.96	0.14	218
	Male	0.99	0.05	257		Male	0.95	0.11	257		Male	0.93	0.15	255
	Total	0.99	0.05	479		Total	0.95	0.10	478		Total	0.94	0.14	473
Computer	Female	0.99	0.03	226	Computer	Female	0.97	0.07	226	Computer	Female	0.96	0.13	222
	Male	0.98	0.08	245		Male	0.92	0.14	243		Male	0.92	0.17	242
	Total	0.99	0.06	471		Total	0.95	0.11	469		Total	0.94	0.16	464
Total	Female	0.99	0.04	448	Total	Female	0.97	0.07	226	Total	Female	0.96	0.13	440
	Male	0.98	0.06	502		Male	0.92	0.14	243		Male	0.92	0.16	497
	Total	0.99	0.05	950		Total	0.95	0.11	947		Total	0.94	0.15	937

Table 4. Mean RTE by Device Condition and Ethnicity

Reading					Science					Mathematics				
Device Condition	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**
Tablet	Black	0.99	0.03	105	Tablet	Black	0.93	0.13	105	Tablet	Black	0.92	0.16	105
	Other	1.00	0.01	64		Other	0.96	0.09	64		Other	0.93	0.16	62
	White	0.99	0.06	310		White	0.96	0.10	309		White	0.95	0.13	306
	Total	0.99	0.05	479		Total	0.95	0.11	478		Total	0.94	0.14	473
Computer	Black	0.98	0.05	87	Computer	Black	0.94	0.12	87	Computer	Black	0.95	0.12	86
	Other	0.99	0.07	79		Other	0.96	0.09	78		Other	0.93	0.17	78
	White	0.99	0.06	305		White	0.95	0.12	304		White	0.94	0.16	300
	Total	0.99	0.06	471		Total	0.95	0.11	469		Total	0.94	0.16	464
Total	Black	0.99	0.04	192	Total	Black	0.93	0.13	192	Total	Black	0.93	0.14	191
	Other	0.99	0.05	143		Other	0.96	0.09	142		Other	0.93	0.16	140
	White	0.99	0.06	615		White	0.95	0.11	613		White	0.94	0.15	606
	Total	0.99	0.05	950		Total	0.95	0.11	947		Total	0.94	0.15	937

-

Table 5. RTE by Item Type

Content Area (# of Items)	Device Condition	N	Mean	SD	Statistical Test	Effect Size (Cohen's d)
Multiple Choice (26)	Tablet	473	0.94	0.11	t =0.94 *p* = 0.35	0.06
	Computer	464	0.94	0.12		
Drag and Drop (11)	Tablet	473	0.99	0.05	t = -0.56 *p* = 0.57	-0.03
	Computer	464	0.99	0.03		
Hot Spot (6)	Tablet	478	0.99	0.06	t = 0.32 *p* = 0.75	0.02
	Computer	469	0.99	0.06		
Fill In the Blank (6)	Tablet	473	0.98	0.09	t = 1.16 *p* = 0.25	0.08
	Computer	464	0.97	0.11		
Multiple Select (6)	Tablet	473	0.94	0.14	t = -0.03 *p* = 0.97	-0.01
	Computer	464	0.94	0.14		
Inline Choice (3)	Tablet	478	0.99	0.06	t = 0.53 *p* = 0.60	0.03
	Computer	469	0.99	0.08		
Graph Point (1)	Tablet	473	0.96	0.20	t = 1.66 *p* = 0.10	0.10
	Computer	464	0.93	0.25		

Table 6. Mean Response Times (in Seconds) by Device Condition and Gender

Reading					Science					Mathematics				
Device Condition	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**
Tablet	Female	1170	429	220	Tablet	Female	1036	315	210	Tablet	Female	667	250	204
	Male	1108	355	255		Male	1067	373	239		Male	582	200	227
	Total	1137	392	475		Total	1052	347	449		Total	622	228	431
Computer	Female	1085	379	226	Computer	Female	995	362	221	Computer	Female	606	225	211
	Male	1074	484	241		Male	910	403	227		Male	535	253	217
	Total	1079	436	467		Total	952	385	448		Total	570	242	428
Total	Female	1127	406	446	Total	Female	1015	340	431	Total	Female	636	239	415
	Male	1091	422	496		Male	990	396	466		Male	559	228	444
	Total	1108	415	942		Total	1002	370	897		Total	596	237	859

Table 7. Mean Response Times (in Seconds) by Device Condition and Ethnicity

Reading					Science					Mathematics				
Device Condition	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**
Tablet	Black	1299	451	105	Tablet	Black	1040	355	96	Tablet	Black	587	225	93
	Other	1235	411	64		Other	1118	367	60		Other	633	234	55
	White	1060	343	306		White	1043	340	293		White	632	228	283
	Total	1137	392	475		Total	1052	347	449		Total	622	228	431
Computer	Black	1265	429	85	Computer	Black	967	395	83	Computer	Black	557	280	82
	Other	1143	406	78		Other	1034	402	76		Other	559	231	72
	White	1011	429	304		White	926	376	289		White	577	233	274
	Total	1079	436	467		Total	952	385	448		Total	570	242	428
Total	Black	1284	440	190	Total	Black	1006	375	179	Total	Black	573	252	175
	Other	1184	409	142		Other	1071	388	136		Other	591	234	127
	White	1036	388	610		White	985	363	582		White	605	323	557
	Total	1108	415	942		Total	1002	370	897		Total	596	237	859

Table 8. Response Time (in seconds) by Item Type

Content Area (# of Items)	Device Condition	N	Mean	SD	Statistical Test	Effect Size (Cohen's d)
Multiple Choice (26)	Tablet	439	1166.56	336.61	**t =3.66 *p* <0.01**	0.25
	Computer	429	1076.09	389.45		
Drag and Drop (11)	Tablet	469	652.64	195.50	t = 1.18 *p* = 0.24	0.08
	Computer	461	636.20	229.36		
Hot Spot (6)	Tablet	468	258.06	88.46	**t = 4.40 *p* <0.01**	0.29
	Computer	461	233.35	82.42		
Fill In the Blank (6)	Tablet	453	240.81	91.82	**t = 3.55 *p* <0.01**	0.24
	Computer	439	219.29	89.14		
Multiple Select (6)	Tablet	423	259.38	107.68	**t = 2.87 *p* <0.01**	0.20
	Computer	423	238.21	107.02		
Inline Choice (3)	Tablet	467	180.43	85.72	**t = 4.21 *p* <0.01**	0.28
	Computer	458	157.38	80.50		
Graph Point (1)	Tablet	453	31.53	19.62	**t = 2.77 *p* <0.01**	0.19
	Computer	433	28.09	17.24		