Device Comparability: Score Range and Subgroup Analyses

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

Laurie Laughlin Davis, Ph.D., Pearson

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

Yuanyuan McBride, Ph.D., Pearson

Xiaojing Kong, Ph.D., Pearson

April, 2016

Abstract

In 2015, the use of tablets for large scale testing programs transitioned from theoretical to reality for several state testing programs. Tablet use for testing is expected to continue to grow over the next several years. The current study evaluated score comparability between tablets and computers across a range of score points, as well as looked at effects for student subgroups (e.g., gender and ethnicity) using data from prior research with high school students testing in reading, math, and science. Results indicated no significant differences between tablets and computers for math and science at any point in the score point range or for any student subgroup. For reading, a small device effect favouring tablets was found for the middle to lower part of the score distribution. This effect seemed to be driven by improvements in performance for male students when testing on tablets. Results are consistent with the growing body of research that suggests minimal impact of device type on student performance.

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

Device Comparability: Score Range and Subgroup Analyses

With the operational launch of the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter-Balanced (SBAC) assessments in 2015, schools across multiple states had the opportunity to deliver large scale assessments to students on tablet devices (PARCC, 2013; SBAC, 2013). These consortia states, as well as others (e.g. Minnesota; MDE, 2015), will soon be joined by the National Assessment of Educational Progress (NAEP; McCullough, 2015) in embracing the use of tablets for digital assessment delivery. Internationally, the Australian Curriculum, Assessment, and Reporting Authority (ACARA) has also announced that it intends to include tablet devices as part of its transition of the National Assessment Program in Literacy and Numeracy (NAPLAN) from paper to digital assessment in 2017[1]. This marks an important transition point for the use of technology in what has historically been called *computer-based testing* (CBT) in that digital delivery of large scale assessment is now inclusive of a wider range of devices than has been available previously.

Research evaluating the use of tablets for assessment has evolved over the past several years from initial small-scale, qualitative research (usability studies and cognitive labs; e.g. Pisacreta, 2013; Strain-Seymour, Craft, Davis, & Elbom, 2013; Davis, Strain-Seymour, & Gay, 2013; Yu, Lorie, & Sewall, 2014) to large-scale quantitative score comparability studies (e.g. Olsen, 2014; Keng, Davis, McBride, & Glaze, 2015; Davis, Orr, Kong, & Lin, 2015; Davis, Kong, McBride, & Morrison, in press). While the qualitative studies have revealed key 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.), these

[1] http://www.nap.edu.au/online-assessment/research-and-development/research-and-development.html

differences have largely not translated into performance differences which impact students' scores.

While the formal psychometric study of device effects across computers and tablets is still nascent, there seems to be a general trend toward non-statistically significant and/or non-practically significant effects at the total score level. For example, in a study of device comparability for the PARCC assessments for students in various grades (i.e., 4, 8, 10), Keng, Davis, McBride, & Glaze (2015) concluded that "most differences between the tablet and computer conditions were small in magnitude and nonsignificant based on the statistical criteria used in the study" (p. 28). While the authors report statistically significant differences in favor of computer for an integrated reading-writing assessment of English language arts for 4th grade students, they attribute this to lack of student familiarity with innovative assessment tasks as well as methodological challenges with creating a matched sample. Similarly, they attribute small effects observed for 4th grade students in mathematics to a lack of familiarity and comfort with entering responses to mathematics tasks on tablets.

These findings are consistent with conclusions from Olsen (2014) who stated that "there is strong evidence that STAR Reading Enterprise and STAR Math Enterprise were measuring the same attribute regardless of device type" (p. 2). This research reported statistically significant, albeit very small effect sizes, for reading. These results varied in direction across grade levels for students in grades 1-11 (i.e., tablet favored in half the cases and computer favored in the other half). The author recommended against making an adjustment to scoring based upon device for reading. The research also found practically small but statistically significant results for math favoring the computer at several grade spans encompassing grades 1-11. The author noted that effect sizes were not large (double the size

of reading effect sizes, but still “small” or “very small”) and concluded that score adjustments were not warranted. However, he did suggest continued monitoring of score differences during operational implementation.

This is also consistent with research from Davis, Orr, Kong, & Lin (2015) which compared computers and tablets (with and without external keyboards) for assessment of written composition and found “no significant performance differences observed in student writing across study conditions” (p. 191). Davis, Kong, McBride, & Morrison (in press) reported no significant differences in mean performance between tablet and computer conditions within content area (reading, math, or science) or for item type across content areas. This research also evaluated the structural differences in test scores between tablet and computer conditions through a measurement equivalence study and found support for both configural and weak measurement invariance. They concluded that both the structure of the test as well as the item loadings were the same across device conditions

Given that the current generation of touch-screen tablets have only been available since 2010, when Apple first launched the iPad, research into the area of device comparability with regard to tablets is a recent development and, to date, has largely been focused on overall group device effects rather than on student sub-groups. Some research has supported the development of assessment applications specifically for special populations which make use of tablet technology (Lopez & Wolf, 2013; Hildago & Sato, 2014). However, the relationship between student group membership and score comparability across computer and tablet devices is a very important, yet unexplored, domain. While aggregated performance largely shows little cause for concern, there may be performance differences for specific student groups that should be identified and addressed to promote equity in large scale assessment.

Additionally, while the Davis, Kong, McBride, & Morrison (in press) study evaluated differences in mean performance between computers and tablets, the research stopped short of a more thorough evaluation which might reveal differences at discrete points in the score distribution that are masked by only looking at the means. The current study conducts a secondary analysis using data from Davis, Kong, McBride, & Morrison (in press) to evaluate device comparability across a range of student ability levels as well as demographic variables for gender and ethnicity.

Method

Two different sets of analyses were conducted to explore differences between computer and tablet conditions in terms of differential effects by score range and student subgroup. The first set of analyses was intended to look at performance differences between device conditions across a range of score points rather than just comparing mean performance differences. Data were analyzed for each of three content areas (i.e., reading, mathematics, and science) using a variant of the Matched Samples Comparability Analysis (MSCA; Way, Davis, & Fitzpatrick, 2006). Five hundred bootstrap replications of a raw score to raw score linking between devices were conducted for each content area. The second set of analyses was designed to look at device effect by subgroup for gender and ethnicity. 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 interactions between device condition and student subgroup.

Participants

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.

Data Analysis

Score Range Analyses

A raw score to raw score equating was conducted between tablet and computer conditions for each content area using a process similar to the Matched Samples Comparability Analysis (MSCA; Way, Davis, & Fitzpatrick, 2006). Student responses from the computer condition and tablet condition were separately calibrated with the Rasch Partial Credit Model through Winsteps using the option to center each calibration on persons (rather than the default to center on items). An expected tablet raw score was calculated for each

student in the computer condition by summing the probabilities of correctly answering each item given estimates of item difficulty obtained from a calibration of student responses in the tablet condition. Assuming no device effect, the expected tablet raw score calculated in this way would be equivalent to the raw score obtained by students of the same ability level in the tablet condition. To the extent that the expected tablet raw score for students in the computer condition is higher than the raw score of students in the tablet condition, this indicates a device effect in favor of tablet (e.g., indicating that it was easier to take the test on tablet than on computer). To the extent that the expected tablet raw score for students in the computer condition is lower than the raw score for students in the tablet condition, this indicates a device effect in favor of computer (e.g., indicating that it was easier to take the test on computer than on tablet). Raw score differences were judged to be practically significant when they exceeded half a raw score point such that student scores would have rounded to the next raw score value ("differences that matter"; Dorans & Feigenbaum, 1994).

This method allows evaluation of device effects at each raw score point in the test and gives a more nuanced understanding of the effects than a simple mean comparison. In MSCA, this raw score to raw score equating is typically conducted across a large number of bootstrapped replications with matched samples being created for the tablet and computer conditions. In this case, because the samples were already assumed to be randomly equivalent, matching variables were not used and five hundred bootstrap samples were drawn randomly from the original sample in order to generate an estimate of the standard error of linking across replications. The mean difference in theta between students in the tablet and computer conditions across replications was computed for each raw score point and compared to the

standard error of linking. Differences greater than two standard errors of the linking were judged to be statistically significant.

Subgroup Analyses

Two sets of two-way analysis of variances (ANOVAs) were conducted for each content area with device condition as one independent variable. The other independent variable was either gender (i.e., Set 1) or ethnicity (i.e., Set 2). Main effects for student subgroup and device condition as well as interaction of device condition, with the student subgroup, were evaluated. Effect size (eta-squared) was calculated for any statistically significant results relating to device condition or interaction of device condition with student subgroup.

Results

Score Range Analyses

Tables 3-5 show the results of the MSCA analyses for reading, math, and science. The columns in the tables are defined as follows:

- RAW –Raw score for students in the tablet condition.
- COMP_RS –Equivalent raw scores on the computer test based on the comparability linking. In other words, this is the raw score of the computer group had they taken the tablet condition. Note that a higher equivalent raw score indicates that the computer version of the test was more difficult.
- RS_SD – Standard deviation of the equivalent raw scores over the replications for the computer condition on tablet. In other words, this is the standard deviation of the raw scores of the computer group had they taken the tablet condition.

- COMP_THETA – Average theta values for the computer group at each equivalent raw score across all replications.
- TAB_THETA – Theta values obtained from the initial Winsteps calibration of the tablet condition in the first iteration.
- THETA_SD – Standard deviation theta values for the computer group at each equivalent raw score across all replications.
- RS_DIF – Difference between raw score equivalent and obtained raw score.
- THETA_DIF – Difference between theta equivalent and obtained theta.
- SIG – Raw score points where scale score differences exceed two standard errors of the linking and where the difference in raw scores is greater than half a point are noted by "*".

In these tables, the equating conversions for the computer and tablet conditions are assumed to be the same for zero and perfect scores, since true score equating conversions cannot be estimated with the Rasch model at these score points. For reading (i.e., Table 3), the differences in the raw score conversions were greater than one-half of a point for half of the obtainable raw score values (raw score values ranging from 5 to 21). In terms of theta, the differences exceeded two standard errors of the linking across the same general raw score range (raw score values ranging from 2 to 21). For math (i.e, Table 4), the differences in the raw score conversions were greater than one-half of a point for a small range of raw score values (raw score values ranging from 14 to 17). However, the differences in theta were less than two standard errors of the linking across the entire score range indicating that differences were not statistically significant. For science (i.e., Table 5), the differences in the raw score

conversions were never as much as one-half of a point. In terms of theta differences, the differences were less than two standard errors of the linking across the entire score range.

Figures 1 through 3 present theta differences between the tablet and computer conditions graphically as a function of raw score, along with upper and lower intervals defined by plus and minus two bootstrap standard errors of equating. These graphs provide a relatively concise summary of the patterns found in Tables 3 through 5.

Subgroup Analyses

ANOVA

Table 6 shows the mean scores for students in each device condition by gender for reading, math, and science, respectively. For reading, neither the main effect for device condition, $F(1, 896)=3.132$, $p>.01$, nor the main effect for gender, $F(1, 896)=2.655$, $p>.01$, were statistically significant. However, there was a significant interaction between device condition and gender, $F(1,896)=10.184$, $p<.01$, with male students in the tablet condition performing higher than male students in the computer condition. The eta-squared value for this interaction was 0.01, indicating a small effect size. Figure 4 shows this interaction graphically. For math, there were no statistically significant differences for either main effect ($F(1, 896)=0.007$, $p>.01$, for device condition; $F(1,896)=0.710$, $p>.01$ for gender) or the interaction between device condition and gender, $F(1,896)=.001$, $p>01$. Similarly, for science, there were no statistically significant differences for either main effect ($F(1,896)=0.152$, $p>.01$ for device condition; $F(1,896)=0.002$, $p>.01$ for gender) or the interaction between device condition and gender, $F(1,896)=1.170$, $p>01$.

Tables 7 shows the mean scores for students in each device condition by ethnicity for reading, math, and science, respectively. For all three content areas, statistically significant

differences were found for the main effect of ethnicity (reading: $F(3,892)=7.203$, $p<.01$; math: $F(3,892)=29.266$, $p<.01$; science: $F(3,892)=21.580$, $p<.01$), but there were no statistically significant differences found either for the main effect of device condition (reading: $F(1,892)=0.345$, $p>.01$; math: $F(1,892)=0.004$, $p>.01$; science: $F(1,892)=0.088$, $p>.01$) or for the interaction of device condition with ethnicity (reading: $F(3,892)=0.894$, $p>,01$; math: $F(3,892)=0.055$, $p>.01$; science: $F(3,892)=0.180$, $p>.01$). As the purpose of this research is to understand the effect of device on subgroups rather than to evaluate performance differences within the subgroups themselves, the significant differences for ethnicity will not be further interpreted.

Discussion

The results of this study suggest that the comparable performance across device conditions seen in the mean comparisons conducted by Davis, Kong, McBride, and Morrison (in press) generally holds across the full range of scores and across student subgroups. The one exception to this was for the reading content domain, where both the score range analysis and sub-group analysis for gender indicated that student performance (specifically for male students) was improved by use of tablets. The direction of this effect may seem surprising when considering the impact of tablet screen size on the overall management of screen real estate for reading assessments. Students must balance the presentation of the reading passage text along with the actual item stimulus and responses in a smaller screen space than with a laptop or desktop computer. With tablet screens, additional scrolling may be needed in order to view the full reading passage and scrolling has fairly consistently been identified as a factor in mode effects between computer and paper test administrations (Way, Davis, Keng, & Strain-Seymour, 2015). Additionally, ergonomic factors such as eye strain and fatiguing

effects resulting from the screen glare might be a specific concern for viewing reading passages on tablets (Davis, Strain-Seymour, & Gay, 2013).

However, Davis, Kong, McBride, and Morrison (in press) did find some evidence in student survey results to suggest that reading may be an area where students prefer to work with tablets. Specifically, they noted that the scrolling interface used to present the reading passages in their study provided 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 through the passage on the computer may be somewhat more cumbersome. The authors additionally cited the increasing use of eReaders and reading applications for hand-held devices that might make reading on touch screen tablets a familiar experience to students.

Why device effects were observed for males in reading, but not females, is less clear. A review of student survey responses did not reveal any differential use in devices between genders which would offer an explanation in terms of either experience level with devices or novelty of devices. One hypothesis is that male students found tablets more engaging than computers and, therefore, were more motivated to provide a stronger level of effort which improved their overall performance relative to the computer condition. However, this effect might have been expected across content areas and was only observed with reading.

It should be noted, however, that the device effect for reading observed in the score range analysis was relatively modest with differences not exceeding a full point at any point in the raw score range for the score range analysis**.** Additionally, the effect size for the interaction between gender and device for reading was small and the sample size for the study was not intended to detect small effect sizes (especially when the sample was split into smaller units for subgroup analysis). As such, this finding should be interpreted with caution and further

research to specifically evaluate any hypotheses of gender by device interactions should be conducted.

While this study raises some interesting questions with regard to the assessment of reading on tablets, the results generally support the continued use of tablet devices in large scale assessment with at least some reasonable expectation of score comparability. While each assessment program should explore potential device effects within its own item and test specifications and testing interfaces, the inclusion of tablets within the allowable group of devices for CBT seems well supported. Certainly, the results for the reading content area suggest that tablets may even offer a better experience for some groups of students. As with all technology, students should have an opportunity to work with tablets as part of their daily classroom activities before they use them in a testing context. Tablets should not be used merely as a "testing device" to extend computer resources on testing day, but should be incorporated into a richer program of technology use within schools' academic programs.

References

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

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

Hildaldo, P., & Sato, E. (2014). *New technologies to assess English learners.* Presented at the annual meeting of the California Educational Research Association. Retrieved from http://images.pearsonassessments.com/images/assets/tell/CERA-Paper-New-Technologies-to-Assess-English-Learners.pdf

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.

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, Annaheim, CA.

McCullough, J (2015). *Delivering the national assessment on tablet: Psychometric challenges and opportunities.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

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.

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

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.

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.

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%)
Hispanic	28 (5.8%)	33 (6.9%)
Other	36 (7.4%)	46 (9.6%)
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%**

Table 3. *Matched Samples Comparability Results for Reading.*

RAW	COMP_RS	RS_SD	COMP_THETA	TAB_THETA	THETA_SD	RS_DIF	THETA_DIF	SIG
0	0.337	0.035	-5.151	-5.582	0.045	0.337	0.000	
1	1.129	0.084	-4.199	-4.301	0.068	0.129	0.102	
2	2.262	0.119	-3.370	-3.502	0.060	0.262	0.132	
3	3.380	0.141	-2.856	-2.999	0.053	0.380	0.143	
4	4.479	0.159	-2.477	-2.623	0.048	0.479	0.146	
5	5.561	0.177	-2.175	-2.319	0.045	0.561	0.144	*
6	6.628	0.193	-1.924	-2.063	0.043	0.628	0.139	*
7	7.680	0.208	-1.707	-1.841	0.041	0.680	0.134	*
8	8.719	0.221	-1.517	-1.644	0.039	0.719	0.127	*
9	9.743	0.233	-1.345	-1.467	0.038	0.743	0.122	*
10	10.762	0.242	-1.188	-1.303	0.036	0.762	0.115	*
11	11.775	0.250	-1.042	-1.152	0.035	0.775	0.110	*
12	12.772	0.256	-0.905	-1.010	0.035	0.772	0.105	*
13	13.762	0.261	-0.776	-0.874	0.033	0.762	0.098	*
14	14.749	0.264	-0.651	-0.745	0.033	0.749	0.094	*
15	15.731	0.266	-0.531	-0.620	0.032	0.731	0.089	*
16	16.708	0.266	-0.414	-0.498	0.032	0.708	0.084	*
17	17.681	0.265	-0.299	-0.379	0.031	0.681	0.080	*
18	18.651	0.263	-0.185	-0.261	0.031	0.651	0.076	*
19	19.610	0.260	-0.072	-0.144	0.031	0.610	0.072	*
20	20.568	0.255	0.042	-0.026	0.031	0.568	0.068	*
21	21.526	0.249	0.158	0.094	0.030	0.526	0.064	*
22	22.475	0.241	0.276	0.216	0.030	0.475	0.060	
23	23.428	0.232	0.399	0.342	0.031	0.428	0.057	
24	24.375	0.222	0.526	0.474	0.031	0.375	0.052	
25	25.316	0.210	0.660	0.613	0.031	0.316	0.047	
26	26.257	0.196	0.804	0.762	0.032	0.257	0.042	
27	27.200	0.182	0.962	0.925	0.033	0.200	0.037	
28	28.141	0.165	1.138	1.108	0.034	0.141	0.030	
29	29.083	0.148	1.339	1.317	0.035	0.083	0.022	
30	30.031	0.128	1.578	1.566	0.037	0.031	0.012	
31	30.988	0.106	1.879	1.879	0.039	-0.012	0.000	
32	31.960	0.081	2.298	2.311	0.042	-0.040	-0.013	
33	32.957	0.049	3.003	3.031	0.04	-0.043	-0.028	
34	33.680	0.018	3.864	4.255	0.022	-0.320	0.000	

Table 4. *Matched Samples Comparability Results for Math.*

RAW	COMP_RS	RS_SD	COMP_THETA	TAB_THETA	THETA_SD	RS_DIF	THETA_DIF	SIG
0	0.312	0.022	-3.675	-4.071	0.028	0.312	0.000	
1	1.034	0.067	-2.779	-2.800	0.062	0.034	0.021	
2	2.060	0.120	-1.983	-2.007	0.070	0.060	0.024	
3	3.085	0.162	-1.470	-1.500	0.068	0.085	0.030	
4	4.109	0.197	-1.076	-1.110	0.067	0.109	0.034	
5	5.132	0.225	-0.747	-0.783	0.067	0.132	0.036	
6	6.160	0.249	-0.456	-0.496	0.066	0.160	0.040	
7	7.193	0.270	-0.193	-0.238	0.065	0.193	0.045	
8	8.231	0.289	0.049	-0.001	0.065	0.231	0.050	
9	9.272	0.306	0.276	0.220	0.064	0.272	0.056	
10	10.322	0.320	0.492	0.428	0.064	0.322	0.064	
11	11.374	0.330	0.699	0.627	0.064	0.374	0.072	
12	12.429	0.337	0.901	0.820	0.064	0.429	0.081	
13	13.480	0.339	1.100	1.009	0.064	0.480	0.091	
14	14.520	0.336	1.298	1.198	0.065	0.520	0.100	
15	15.555	0.327	1.501	1.390	0.065	0.555	0.111	
16	16.563	0.314	1.708	1.589	0.066	0.563	0.119	
17	17.550	0.295	1.926	1.800	0.068	0.550	0.126	
18	18.508	0.271	2.159	2.028	0.070	0.508	0.131	
19	19.432	0.240	2.419	2.286	0.073	0.432	0.133	
20	20.322	0.201	2.722	2.592	0.079	0.322	0.130	
21	21.191	0.151	3.121	2.994	0.094	0.191	0.127	
22	22.059	0.088	3.723	3.648	0.088	0.059	0.075	
23	22.701	0.031	4.447	4.787	0.035	-0.299	0.000	

Table 5. *Matched Samples Comparability Results for Science.*

RAW	COMP_RS	RS_SD	COMP_THETA	TAB_THETA	THETA_SD	RS_DIF	THETA_DIF	SIG
0	0.285	0.020	-4.444	-4.811	0.025	0.285	0.000	
1	0.952	0.055	-3.587	-3.522	0.066	-0.048	-0.065	
2	1.914	0.091	-2.791	-2.719	0.069	-0.086	-0.072	
3	2.885	0.119	-2.278	-2.219	0.058	-0.115	-0.059	
4	3.859	0.145	-1.901	-1.848	0.052	-0.141	-0.053	
5	4.839	0.168	-1.597	-1.548	0.049	-0.161	-0.049	
6	5.824	0.188	-1.337	-1.291	0.047	-0.176	-0.046	
7	6.819	0.203	-1.102	-1.060	0.046	-0.181	-0.042	
8	7.826	0.214	-0.884	-0.846	0.045	-0.174	-0.038	
9	8.841	0.221	-0.674	-0.641	0.045	-0.159	-0.033	
10	9.873	0.225	-0.468	-0.443	0.044	-0.127	-0.025	
11	10.908	0.227	-0.267	-0.249	0.044	-0.092	-0.018	
12	11.950	0.229	-0.069	-0.059	0.043	-0.050	-0.010	
13	12.991	0.231	0.127	0.129	0.043	-0.009	-0.002	
14	14.036	0.233	0.321	0.314	0.043	0.036	0.007	
15	15.080	0.235	0.512	0.498	0.043	0.080	0.014	
16	16.118	0.237	0.700	0.679	0.043	0.118	0.021	
17	17.148	0.238	0.888	0.861	0.043	0.148	0.027	
18	18.172	0.237	1.075	1.043	0.044	0.172	0.032	
19	19.181	0.233	1.264	1.229	0.045	0.181	0.035	
20	20.169	0.226	1.458	1.423	0.046	0.169	0.035	
21	21.151	0.214	1.662	1.628	0.047	0.151	0.034	
22	22.115	0.200	1.882	1.852	0.049	0.115	0.030	
23	23.075	0.183	2.128	2.104	0.051	0.075	0.024	
24	24.029	0.165	2.413	2.399	0.055	0.029	0.014	
25	24.987	0.144	2.763	2.762	0.060	-0.013	0.001	
26	25.956	0.117	3.232	3.245	0.067	-0.044	-0.013	
27	26.952	0.075	3.987	4.020	0.065	-0.048	-0.033	
28	27.677	0.028	4.875	5.282	0.036	-0.323	0.000	

Table 6. *Mean Scores by Device Condition and Gender*.

Reading					Math					Science				
Device Condition	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**	**Device Condition**	**Gender**	**Mean**	**SD**	**N**
Tablet	Female	19.39	4.645	215	Tablet	Female	8.83	5.519	215	Tablet	Female	12.65	4.285	215
	Male	19.87	4.857	217		Male	8.87	5.143	217		Male	12.86	4.048	217
	Total	19.63	4.753	432		Total	8.85	5.327	432		Total	12.75	4.164	432
Computer	Female	19.92	5.238	245	Computer	Female	9.16	5.871	245	Computer	Female	12.96	5.057	245
	Male	18.24	5.430	223		Male	9.17	5.840	223		Male	12.52	4.702	223
	Total	19.12	5.390	468		Total	9.16	5.850	468		Total	12.75	4.891	468
Total	Female	19.67	4.971	460	Total	Female	9.00	5.705	460	Total	Female	12.82	4.709	460
	Male	19.05	5.214	440		Male	9.03	5.503	440		Male	12.68	4.390	440
	Total	19.37	5.098	900		Total	9.01	5.604	900		Total	12.75	4.554	900

Table 7. *Mean Scores by Device Condition and Ethnicity.*

Reading					Math					Science				
Device Condition	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**	**Device Condition**	**Ethnicity**	**Mean**	**SD**	**N**
Tablet	Black	18.13	4.867	100	Tablet	Black	6.03	4.356	100	Tablet	Black	10.65	3.971	100
	Hispanic	18.63	4.217	27		Hispanic	6.85	5.082	27		Hispanic	12.52	3.926	27
	Other	20.41	5.129	34		Other	11.03	5.579	34		Other	13.85	5.439	34
	White	20.20	4.949	299		White	9.96	5.776	299		White	13.45	4.711	299
	Total	19.67	4.971	460		Total	9.00	5.705	460		Total	12.82	4.709	460
Computer	Black	18.07	5.331	82	Computer	Black	6.09	3.923	82	Computer	Black	10.32	3.893	82
	Hispanic	17.97	5.245	32		Hispanic	7.22	5.11	32		Hispanic	12.13	3.916	32
	Other	21.02	4.826	42		Other	10.90	6.092	42		Other	14.33	4.124	42
	White	19.15	5.162	284		White	9.80	5.507	284		White	13.19	4.369	284
	Total	19.05	5.214	440		Total	9.03	5.503	440		Total	12.68	4.39	440
Total	Black	18.10	5.067	182	Total	Black	6.05	4.155	182	Total	Black	10.50	3.929	182
	Hispanic	18.27	4.774	59		Hispanic	7.05	5.056	59		Hispanic	12.31	3.892	59
	Other	20.75	4.940	76		Other	10.96	5.83	76		Other	14.12	4.730	76
	White	19.69	5.077	583		White	9.88	5.642	583		White	13.32	4.546	583
	Total	19.37	5.098	900		Total	9.01	5.604	900		Total	12.75	4.554	900

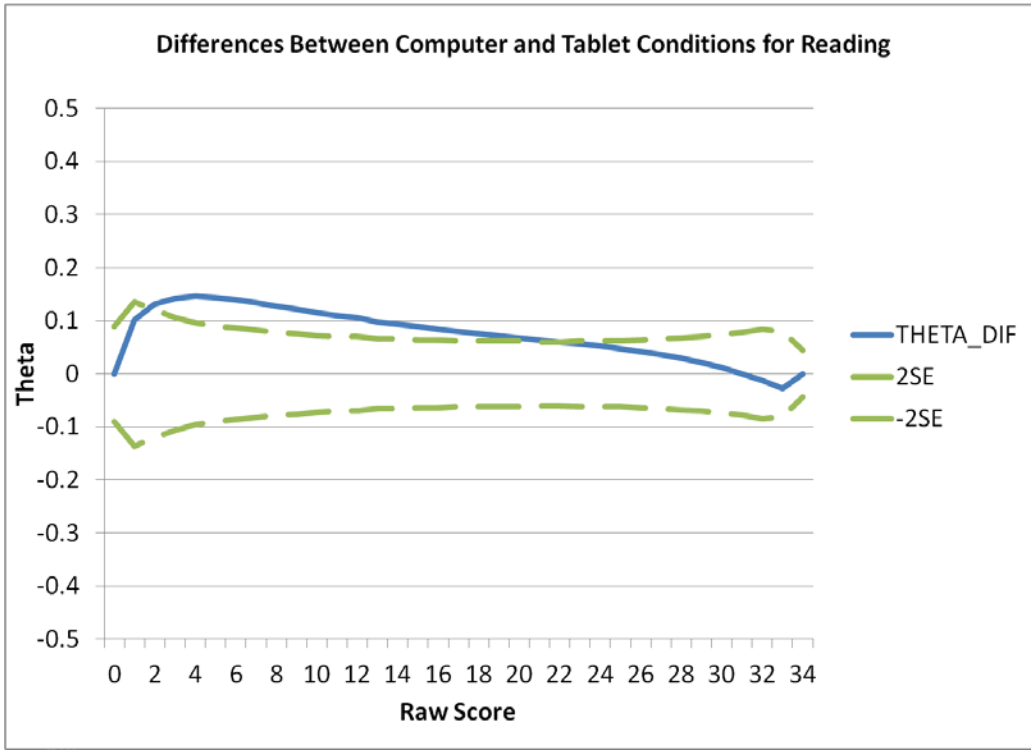

Figure 1.*Theta Differences between Computer and Tablet Conditions for Reading.*

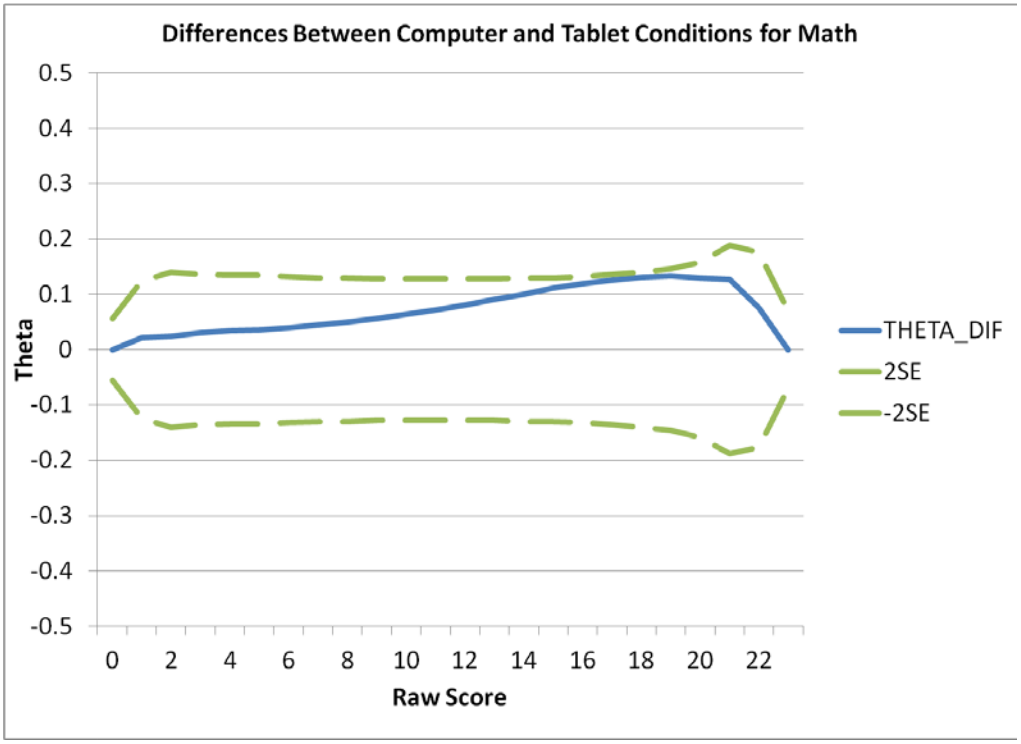

Figure 2.*Theta Differences between Computer and Tablet Conditions for Math.*

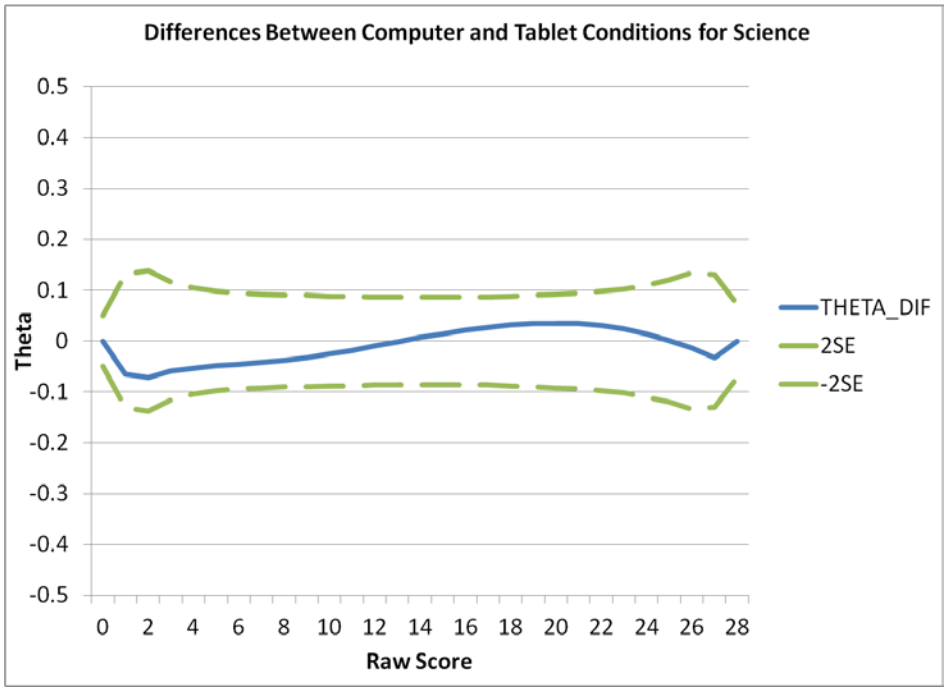

Figure 3.*Theta Differences between Computer and Tablet Conditions for Science.*

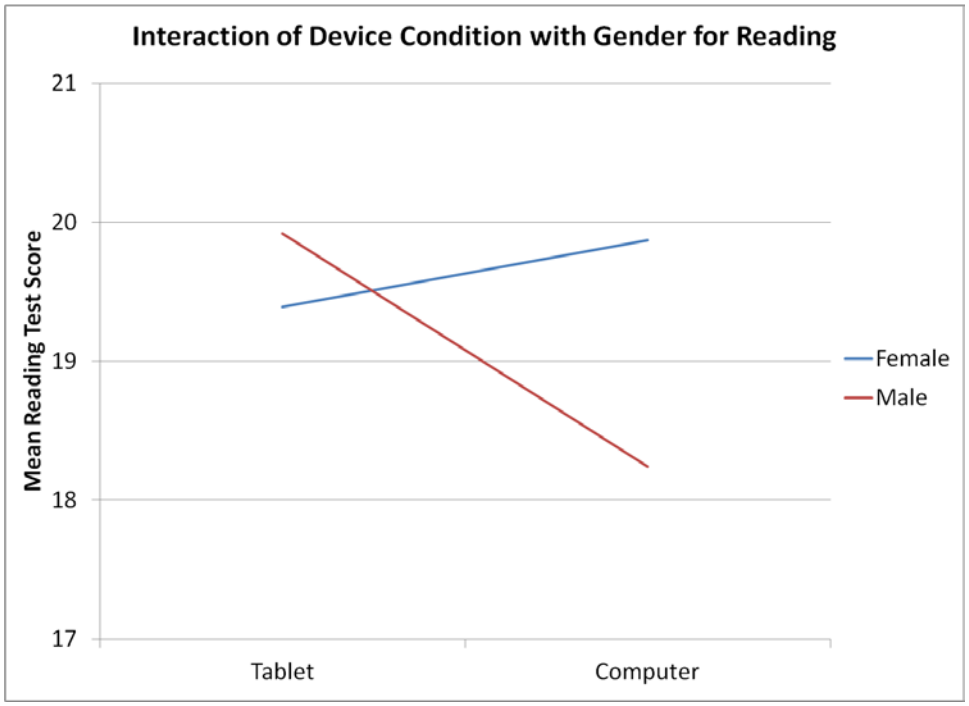

Figure 4. *Interaction of Device Condition with Gender for Reading.*