A Comparison of Multiple-choice and Technology-enhanced Item Types Administered on Computer versus iPad

By

Tonya Eberhart

Submitted to the graduate degree program in Curriculum and Teaching and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Chairperson Dr. Marc Mahlios

_____
Dr. Douglas Huffman

_____
Dr. Neal Kingston

_____
Dr. Phil McKnight

_____
Dr. Steven White

Date Defended: September 8, 2015

The Dissertation Committee for Tonya Eberhart
certifies that this is the approved version of the following dissertation:

A Comparison of Multiple-choice and Technology-enhanced Item Types Administered on Computer versus iPad

_____
Chairperson Dr. Marc Mahlios

Date approved: September 10, 2015

Abstract

It is not known if various technology device differences used for high-stakes assessment result in comparable student performance results. It is important to examine the cognitive processes and motivation of students using various technology devices and item types. The purpose of this study was to compare student performance and explore student motivation when tested using technology-enhanced and multiple-choice items on a computer and a tablet.

This study used a mixed-method based research that included concurrent verbal protocols (Think-alouds), a short, self-report questionnaire, and English language arts (ELA) and mathematics assessment scores. Quantitative and qualitative findings suggest that device-type differences in student performance scores are small, and item-type and differences are moderate to large. Many factors affect student performance on high-stakes assessment and more research is needed to further understand how item-type and device-type affect student performance.

CONTENTS

A Comparison of Multiple-choice and Technology-enhanced Item Types Administered on Computer versus iPad

Chapter 1: Introduction

There are major shifts in the United States education system occurring in the arenas of educational standards, assessment and technologies. Technology devices and infrastructures are being invested in and placed into schools for instructional and assessment purposes at an increasingly rapid rate, with most teachers currently using technology in their classrooms to some degree (Bill & Melinda Gates Foundation, 2012b). Furthermore, the Common Core State Standards (CCSS) initiative is forging towards college- and career-ready standards (CCSS, 2014). Two state consortia – the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC) - were awarded Race to the Top grants to create next-generation assessments aligned to the CCSS (PARCC, 2014; SBAC, 2014). Consequently, various technology devices are being used in the classroom with little empirical evidence of the positive impact on learning, and test developers are experimenting with technology-enhanced item types in K-12 testing presumably to provide better and more efficient testing.

This dissertation used mixed-methods based research to report on student performance and motivation using multiple-choice and technology-enhanced item test types on various technology devices. This study utilized concurrent verbal protocols (Think-alouds), a short, self-report questionnaire, and English language arts (ELA) and mathematics assessment scores. The objective of this research was to determine the comparability of student performance and explore student motivation when tested using multiple-choice and technology-enhanced item types delivered via computer (desktop or laptop) or a tablet (iPad).

Problem Statement

It is not known if various technology device differences used for high-stakes assessment result in comparable student performance results. In addition, it is important to examine the cognitive processes and motivation of students using various technology devices and item types. There is little empirical evidence to support the impact of the use of technology in the classroom for instructional purposes (Bill & Melinda Gates Foundation, 2012). Whether or not the benefits outweigh the concerns about technology use in the classroom for instruction and assessment, the technology is being infused into the classroom at a rapid rate. Schools across the country have been ramping up technology in the classrooms, with iPad/tablets and one-to-one initiatives occurring in many districts. For example, in 2011 Tower School in Marblehead, Massachusetts started a one-to-one program using iPads. The program for third through eighth grade called for all students to have an iPad. Along with the program, wireless infrastructures were upgraded to handle the additional network needs and teachers were given summer training for using iPads. One purpose of the program was to bring their students to the "forefront of the information and technology revolution" (Taborn, 2011). In Maryland in 2014, a school district implemented a bring-your-own-device (BYOD) program in three schools. Gradually, every student in the district will have the ability to utilize their own tablet, laptop or other mobile device in the classroom (Hart, 2015). Example after example can be found in the literature of schools and districts forging ahead in the technology revolution for their students. The Office of Educational Technology as part of the National Education Plan of 2010 discussed that the challenges for our educational leaders is to take advantage of the technology used in our real-world activities to equip our students with stimulating learning tools to engage students with

content, and to use as a resource for assessment that will more accurately and authentically measure student achievement (Atkins et al., 2010). These initiatives continue as educators and policy-makers call for strategies to keep up with emerging innovations in technology.

Traditional large-scale testing has relied heavily on text- and image-based multiple-choice questions. However, in order to measure the deeper knowledge, skills and abilities (KSAs) and higher-order thinking skills, as emphasized within the CCSS Initiative (Common Core State Standards Initiative, 2014), some criticize multiple-choice item types as limited in the depth of knowledge and skills they can assess (Jodoin, 2003). Performance-based tasks improve the ability to assess deeper knowledge and skills, but are disadvantageous in that they may still limit domain coverage (Hambleton, 2000) and cannot be automatically scored by the computer; furthermore, inter-rater scoring reliability is difficult and the cost associated with expert raters in large-scale testing is prohibitive (Clauser, 1997).

One solution to better assess the more in-depth and complex KSAs may come through the use of technology-enhanced item types. It is presumed that technology-enhanced item types provide a more authentic interaction with the construct being assessed, allowing for better and more efficient testing. Technology-enhanced item types allow students to interact with the item, dragging and dropping text or objects, reordering images, graphing number lines, providing visual representations of fractions, and highlighting text within passages. However, there have been few studies of these items with the K – 12 population. Moreover, these items are given on a multitude of devices (tablets, laptops, and desktops) and students interacting with the items have a multitude of backgrounds and experience with the devices. One particular consideration with the technology-enhanced items is the difference between using a mouse to manipulate an item on a desktop/laptop and using a finger on a tablet. Additionally, there is a considerable difference in

screen size. It is import to examine whether any differences in technology devices result in any differences in performance.

The purpose of this study was to compare student performance and explore student motivation when tested using technology-enhanced and multiple-choice items on a computer and a tablet. In particular, when students are tested using different item types and different devices it is important to ensure one group is not advantaged over another because of the device type or the item type, and to explore if student affect or motivation result in differences. Therefore, there are three facets to this study that join together in a substantial way: technology device type comparison, item type comparison, and exploration of student experience with item-types and technology types.

Purpose of the Study

The purpose of this study was to compare student performance and explore student motivation when tested using technology-enhanced and multiple-choice items on a computer and a tablet. There are many differences to consider when using new item-types and new technologies. Do screen size differences give an advantage to one group? Do the manual differences between using a mouse to manipulate an item on a desktop/laptop and using a finger on a tablet result in any difference in student performance? Are technology-enhanced items more engaging, and therefore, more motivating than multiple-choice item test types resulting in performance differences? As the trend in assessment continues to grow towards the use of technology-enhanced item types and tests are being delivered on a variety of devices, it is important to determine if any differences need to be mitigated.

Significance of the Study

This research will contribute to the literature by providing initial evidence of the comparability and usability of technology-enhanced item types and multiple-choice item types, considering the variety of devices being used for test administration. It's important to examine whether there are any item-type or device-type differences in student performance scores and whether student performance is affected by student motivation. It was the ultimate hope that this study will contribute to an approach for best practices for developing families of item types that behave predictably and measure the targeted construct, whether administered on a desktop/laptop or tablet.

Theoretical Framework

Many teachers currently use some form of technology to promote learning in the classroom (Bill & Melinda Gates Foundation, 2012a; Pressey, 2013). Specifically, Apple's tablet, the iPad has been inserted into classrooms faster and more broadly than any other computer device (Vu, McIntyre, & Cepero, 2014). What does this mean for educators who do, or don't, have iPads in their classrooms when it comes to instruction and assessment? Does using technology in the classroom contribute to student engagement and motivation? With many teachers already using tablets for instruction, it's important to determine if students' find using technology for classroom instruction and assessment motivating and engaging. Equally important, is to ensure that using these devices for assessment is comparable to the desktop/laptop administration.

How do test developers ensure that students are being assessed fairly and authentically in the midst of the push for a wider use of technology device types in the administration of large-scale assessment? Technology and computer devices will likely remain a major part of schooling in

the United States. Large investments of money and time are being made to improve quality and cost-effectiveness of assessments (SBAC, 2014). As technology, and as a consequence, assessment items and administration devices continue to advance, research must continue to extend understanding between technology and assessment. Kozma (1991) stated:

> Ultimately, our ability to take advantage of the power of emerging technologies will depend on the creativity of designers, their ability to exploit the capabilities of the media, and our understanding of the relationship between these capabilities and learning. A moratorium on media research would only hurt these prospects. (p. 206)

Research Questions

1. Is there a significant difference in seventh-grade student performance on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop?
2. Is there a significant difference in seventh-grade student performance on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?
3. Can technology-enhanced item types promote motivation to perform better than multiple-choice item types?
4. Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?

Summary

Schools and educators have begun the arduous and costly task of evaluating broadband and educational technology needs in order to support computerized instruction and assessment. As the consortia member states and independent states move toward computerized instruction and testing, the types of items and devices utilized has the potential to have great influence on

instruction and evaluation of our educational system. Advanced technologies have the potential to differentiate learning and assessment, create richer and deeper learning environments and shape formative and summative assessment strategies. Moreover, technology-enhanced assessments provide for the ability to expand the construct and content being assessed (Parshall & Harmes, 2007).

Although the literature is expanding with regard to new technologies in instruction and assessment, caution should be taken as it appears that some educators and test developers alike have embraced these new technologies as a whole as the answer to education reform. There are many pieces that come together in this current education shift, and this research attempts to begin the exploration into student learning and assessment through the use of different technological devices and assessment item types.

Definition of Terms

Achievement Construct

Haladyna & Downing (2004) describe this concept as being of two kinds. The first as a large set of skills and knowledge and the second is a procedural ability, as in problem solving skills.

Construct Irrelevant Variance

When an assessment item introduces extraneous, uncontrolled variables outside of the domain, or achievement construct, this is called “construct irrelevant variance” (Haladyna & Downing, 2004).

Mixed Methods

Mixed Methods research combines qualitative and quantitative research that has a primary philosophy of pragmatism. It is an approach to theoretical and practical knowledge that combines multiple perspectives, viewpoints and positions (Johnson, Onwuegbuzie, & Turner, 2007).

Motivation

- *Extrinsic motivation* is participating in an activity for external rewards (Deci, 1972).
- *Intrinsic motivation* exists in the relation between individuals and activities. It takes place while learners are driven to perform in an interesting activity.

Response Action

Response action refers to both the input device and the motor action required to respond to the question. Typically the input device is the keyboard and/or the mouse. The examinee might be asked to type in a short response using keyboard strokes, or use the mouse to drag and drop elements in order to create the correct answer. They might need to click on an element and click on another element in order to show the relationship between the two, or to show a new location for the image (Parshall & Harmes, 2007).

Scoring Method

The scoring method for technology-enhanced item types is a set of rules, or algorithm, that defines the score that will be produced based on a predefined score response set of data. There are many things to consider when defining a score set. With multiple-choice questions, one point is given for one correct response selected by the examinee. With technology-enhanced item types, it must be considered whether a graphing item will be full credit only for a correct

response, or partial credit will be given for a correct point or slope given other pieces of the answer are incorrect (Measured Progress/ETS Collaborative, 2012; Parshall & Harmes, 2007).

Also to be considered with scoring, is whether it is automated or manual. Automated scoring is any type of item type that can be automatically scored by the computer or online testing system. Manual scoring includes any type of task that is not able to be scored by a computer system and must be human-scored. For example, essay performance tasks must be scored manually.

Technology-Enhanced Template

Templates are the structure of the technology-enhanced items. The template describes the framework for the innovative items and can be used in the development of the individual items. The template informs the layout, screen design, interaction, and response action of the individual item. Templates can be used across all grades and content areas; however, some templates are of better use in a particular grade or subject. For example, a select text template is of better use for English language arts where the examinee would be required to select a word or sentence from a passage. A graphing template is used for mathematics (Measured Progress/ETS Collaborative, 2012; Parshall & Harmes, 2007, 2009; Zenisky & Sireci, 2002). Table 1 presents a description of each of the technology-enhanced item types in this study.

Table 1
Technology-enhanced Template Types

Template Type	*Brief Description*
Drop-Down	Examinee chooses the correct answer from a drop-down list of options.
Drag-and-Drop	Examinee selects and drags a label, an image or text to a predetermined drop-zone in the response area (an image, area of text, or label area).
Graphing	Examinee clicks to add a point to a given grid/graph. An additional question option is to add another point to snap a line in place.
Matching Lines	Given a word(s), sentence, number(s), mathematical expression/equation and/or object, the examinee clicks to match the appropriate corresponding word(s), sentence, number(s), mathematical expression/equation and/or object.
Multiple Columns	Examinee chooses from a given number of options on a matrix. Typically, three or more stems are on the left, and answer choice(s) are given on the right side of the page. Examinee clicks to choose the radio button of the appropriate choice to fit the stem.
Ordering	Examinee orders elements by dragging them into the correct order, for example, chronologically or smallest to largest.
Select Text	Given a sentence or passage, examinee highlights a word(s) or sentence.

Chapter 2: Review of the Literature

Introduction

The purpose of this dissertation was to compare student performance and explore student motivation using technology-enhanced and multiple-choice items on a computer and a tablet for administration of assessments. This review was based on history, theories and research of accountability testing, computer-based testing, and the use of technology in instruction and assessment. More specifically, studies related to assessment include: the history and studies of school reform, comparisons of paper-based and computer-based measurement of student performance, student motivation and engagement, and computer-device differences that relate to instruction and assessment of performance. The remainder of this chapter examines the historical, empirical and theoretical foundations of this study. Previous and current data for this study was gathered by: (1) searching electronic databases (ERIC, ProQuest, Google Scholar, e-Journals), (2) searching relevant published books, and (3) reviewing bibliographies of journal articles. An examination of related literature follows.

The Beginning of Accountability Testing

With the publishing of *A Nation at Risk* (National Commission on Excellence in Education, 1983), the nation's schools and educators were called into question regarding their ability to educate and prepare our young people for post-secondary education. It cautioned that:

> . . . while we can take justifiable pride in what our schools and colleges have historically accomplished and contributed to the United States and the well-being of its people, the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people. What was

> unimaginable a generation ago has begun to occur - others are matching and surpassing our educational attainments. (p. 112)

A Nation at Risk (1983) asserted that our schools were failing by pointing out, among other things, that our nation's students compare poorly to other nation's students, SAT scores have declined dramatically, and that around thirteen percent of students at age 17 could be deemed functional illiterate and that number is more than twice as high for minority students. The report made recommendations across five major areas: Content, Standards and Expectations, Time, Teaching, Leadership and Fiscal Support. Within the Standards and Expectations category, the report recommended higher academic standards and that those expectations be measured yearly by standardized tests of student performance.

In 1965, President Lyndon B. Johnson signed *The Elementary and Secondary Education Act* (ESEA) into law (U.S Department of Education). This law set out to improve the equity of educational opportunities for students by offering federal funds to low-income youth and to improve the overall quality of schools for all students. In 2001, Congress reauthorized ESEA and in 2002, President George W. Bush signed a new bill into law called *The No Child Left Behind Act* (NCLB) ("No Child Left Behind (NCLB)," 2002). This educational reform remarkably increased the federal governments influence on school policy and promised increased accountability and performance of U.S. schools. NCLB called for states to create assessments that measure and monitor student performance against a high set of common academic standards. In the end, NCLB was the driving force in a new era of public accountability that required schools to set high academic standards and measure those standards through assessments.

The Use of Technology in Education

The use of technology in educational settings has increased rapidly in the last two decades. In 2003, the ratio of students to instructional computers with Internet access in public schools was 4.4 to 1, a decrease from the 12.1 to 1 ratio in 1998, when it was first measured (Parsad, Jones, & Greene, 2005). As teachers have greater access and use of technologies for instructional purposes, the use of computer technologies in testing is felicitous and brings with it many advantages, as well as concerns.

Technology has become a fundamental part of society and is becoming an integral part of classroom instruction. In a report from the National Center for Educational Statistics, 85 percent of students in the fourth grade stated they use a computer at home and 69 percent use the internet at home. Eighty-eight percent of students in eighth grade stated they use a computer at home and 79 percent use the internet at home. As for using computers at school, 86 percent of students in fourth grade and 80 percent of students in eighth grade report computer use at school (Sandene et al., 2005). In the workplace, 96 percent of Americans that have jobs use communication technologies in their lives day-to-day, and 62 percent of Americans with jobs report internet use as an important part of their work. These computer and online skills are a necessary piece in finding, applying to, and securing a job (United States Department of Commerce, 2011). Even though technology has quickly renovated our daily lives, there are many presumed benefits of, and arguments against, the use of technology in the classroom.

There are many perceived benefits to the inclusion of technology in educational settings. The use of technology allows teachers the ability to reach students with a variety of learning styles. Many teachers find that the use of technology helps with student motivation since it enables students to learn in a variety of ways. The use of technology in the classroom better

prepares students for the future since nearly all careers utilize some form of technology as part of the job. Technology can also help students to feel that they are more in charge of their own learning by giving the ability to work at their own pace. Students learn the strategy of looking up information, processing information, and evaluating the information. Computers allow for teachers to drill students through the use of games or educational software. For example, many games and programs exist to drill students on multiplication facts. The internet allows students and classrooms to communicate with other students in other parts of the world, directly from the classroom. This allows for the sharing of information, collaborative learning, and sharing of culture in a way never before possible (Honeycutt, 2013). Some proponents of using technology as the guiding source of student education (e-learning) argue that digital media gives students access to a broad range of sources from which to gather information. It is comparable to when the student reaches the workforce and must rely on the ability to find information and use that information to problem-solve (Birch, 2011).

Collins and Halverson (2009) argue that if schooling today is to survive, it must make major revisions. They submit that students must learn to find and research information and use that information to make decisions through information evaluation and synthesis, much like anyone in the work force today must come to spontaneous conclusions using the instantaneous access to information. Rather than thinking of education as a place where the student goes to absorb a knowledge bank, schooling should be a place where the student learns to access information and use that ability to gain knowledge. Education must keep up with society's use of technology and use technology as a means for lifelong learning. Additionally, the authors point out that students engage with technology faster than their parents and teachers, and spend a significant amount of time web-gaming and/or social-networking. This places more pressure on our traditional

schooling methods to keep up with technology or it will seem antiquated and useless to the student. The authors also recognize the valid argument that teachers note – that there is not much empirical evidence to support that technology improves performance of students. They additionally note other often cited obstacles that must be overcome in order for technology to work to the benefit of instruction, such as cost and the time necessary for implementation.

Some educators are apprehensive about technology being used to replace the teacher. Technology does not have the capacity to care about the student and is not able to respond to the individual student with compassion and the knowledge of the student's background in mind. Teachers care about their students and their well-being and education; technology cannot care about the student. Some students don't always give the use of technology in the classroom the warm reception anticipated. Students' expectations about technology and the compatibility with the students' educational values determines the students' intended use of technology. Sometimes students have the expectation to receive instruction from their teacher in order to learn more productively (Chen, 2011).

Teachers report the biggest barrier to using technology in the classroom effectively is that they don't have acceptable access for students. Obstacles include those such as the need to reserve digital media for use in the classroom, limited availability of computers, and the fact that many students don't have access at home which makes it more challenging to use computers for teaching. Teachers note concern that there is little empirical evidence to support the benefits of using technology for instructional purposes, as was found in a recent survey:

> "Teachers are concerned about the true benefits of technology for their students. Despite recognition of the importance of technology in education, many teachers report skepticism about the actual benefits of using technology in the classroom because there is

little, widely accepted proof that technology tools provide real value for student learning" (Bill & Melinda Gates Foundation, 2012, p. 2).

Similarly, many teachers feel as though they are not sure they are using technology as a tool efficiently and appropriately in order to promote student learning, even though they also believe that technology has helped to engage their students (Bill & Melinda Gates Foundation, 2012; Collins & Halverson, 2009, p. 32). To that end, teachers want more professional development to help them learn the best ways to use technology and more planning time.

Whether or not the benefits outweigh the concerns about technology use in the classroom for instruction, the technology is being infused into the classroom at a rapid rate. In August of 2013, Los Angeles Unified School District began the rollout of iPad tablet computers in their plan to get iPads to each student in the district. In this district seventy percent of student population comes from low-income families and the hope is that this will even the playing field for these students who might not have access to this technology at home (Blume, 2013). The Natick public school in Massachusetts bought MacBook laptops for every eighth-grade student, and the second year for every high school student. The superintendent of Natick district noted that prior to the MacBook one-to-one initiative, students were completing their homework at home on a computer, then coming to school and using paper, pencil and a textbook for instruction. The end-goal for the Natick district initiative, simply put, is to increase student achievement. The Auburn, Maine school district experimented with kindergarteners to see if iPads would improve literacy scores. Half of the sixteen kindergarten classes were randomly assigned iPads and the remaining kindergarten classes were given iPads twelve weeks after the first class. The instruction for all classes was similar. According to the district, the students who received the iPads at the beginning performed better on all measurements of literacy. They speculate this is because

students were able to receive immediate feedback and more time was left for working independently. The district then planned to expand the iPad initiative to first and second grade each consecutive fall semester (Fairbanks, 2013).

Many districts are utilizing the "bring your own device" or "bring your own technology" initiative as a cost-effective way to infuse technology into instruction. At Forsyth County schools in Georgia, every day over 11,000 students hook up to the wireless network at the district using their own device. The district representative notes that the advantage to this program is that when the device used by the students varies, it forces teachers and students to personalize the way that the student learns (Fairbanks, 2013).

Paper-based testing (PBT) compared to Computer-based testing (CBT)

With the use of technology in testing, testing administrators are able to devise new and innovative ways to create, administer and score tests. The face of testing has changed remarkably over the past two decades; the vast majority has moved from a paper-based to a computer-based platform. By 2012, thirty-three states had moved at least partially to online computer-based assessment (SETDA, 2015). In many large-scale testing programs, computer-based testing when compared to paper-based testing has the appeal of faster scoring at a lower cost and the ability to test many skills in a short period of time. Computer-based testing has the additional appeal of future capabilities of assessing new skills and abilities that have previously been too costly or impossible to assess. When computer-based testing was introduced more than twenty years ago, many researchers were interested in determining whether computer-based testing was truly analogous to paper-based testing.

Many researchers have studied the comparability of paper-based tests (PBT) to computer-based tests (CBT). Although results seem to vary across studies, at most the differences appear to

be quite small. The same holds true when comparing PBT to CBT across subgroups and content areas. MacCann (2006) did not find any comparability differences with regard to gender in a study on Australian students. He points out in this study that students were self-selected for this study, which might imply that this was not a typical sample. It could be that females who have a higher familiarity with computers volunteered for the study which would skew the results. The study also found that low socioeconomic (SES) groups scored slightly higher on PBT than on CBT. McCann points out that it could be the case that low SES students might have less access to computers and therefore, fewer computer skills and familiarity; additionally, they may have a negative affective response to CBT which would in turn negatively affect test scores.

In the Gallagher et al. (2002) study, they found a small difference for ethnic groups and a difference on some tests for female examinees. African-Americans and to a lesser degree Hispanic examinees appear to benefit slightly from CBT. For white female examinees, Kingston (2009) noted in a synthesis of studies done from 1997 to 2007, there was no comparability differences based on gender from PBT to CBT.

Choi et al. (2013), in a Web-based language test done in Korea, found that PBT to CBT are comparable. They found the content analysis highly comparable; and overall results of construct-related validation studies indicate comparability of the subjects' scores across computer-based testing to paper-based testing platforms. Poggio et al. (2005) found very little difference in CBT to PBT in a comparability study with seventh-grade students. As in the MacCann study, seventh graders in the target population were self-selected. Students were given the CBT test, and later given the PBT test. Additionally, there were no differences found based on gender or SES.

Many studies have been performed to assess whether computer familiarity, computer anxiety and/or attitudes towards computers affect test performance. Odo (2012) investigated the relationship between computer familiarity and CBT performance in a large city in western Canada. The participants, aged 13-19, were given an online test of second language reading and responded to a paper-based questionnaire about their computer familiarity. The results showed a small but significant variability on CBT scores. Taylor et al. (1998) found that computer familiarity does not affect test performance. Taylor's study examines the test performance of 1,204 examinees on the TOEFL (Test of English as a Foreign Language). The participants were given a questionnaire; and then based on the answers given, they were then classified into one of three computer-familiarity groups: low, high, and moderate computer familiarity. A CBT tutorial was administered to participants prior to testing. Examinees were then given the CBT examination with test items similar to the actual computer-based TOEFL. The items on the test required examinees to respond by: (a) clicking on a picture or letter where it should be placed on a diagram, chart, or picture, (b) selecting two answers, (c) matching or ordering information, or (d) clicking on a correct word or phrase. CBT's were taken on a laptop computer. They found that there was not a meaningful relationship when comparing the amount of familiarity with the computer and student performance on the CBT.

History of computer-based testing

Paper-and-pencil tests remain part of educational assessment today, although most schools have transitioned to computer-based testing and next-generation assessments. At the Tenth Annual Maryland Assessment Conference in October 2010, it was stated that in a survey of state testing directors that 44 states currently have computer-based testing initiatives. Out of those 44 states, 26 were administering large-scale testing online and 15 were planning to administer their

large-scale testing online (Martineau & Dean, 2010). By 2012, thirty-three states were using online computer-based assessments to some degree (SETDA, 2015). In 2014, new online, large-scale state assessments were administered widely across the U.S for the first time that included innovative item types. Within the Partnership for Assessment of Readiness for College and Careers consortium, 16 states administered next-generation, large-scale testing online (PARCC, 2014), and 21 states and the U.S. Virgin Islands administered their large-scale, next-generation assessments online under the SBAC consortium (Smarter Balanced Assessment Consortium, 2012). The new assessments created by the consortium employ the use of innovative item types in an effort to create tests that are more authentic and are a better measure of student performance.

Computer-based assessments have traditionally consisted of multiple-choice questions with text and, when appropriate, static images. The examinee response action consisted of selecting a radio button(s) for the answer choice. There are many advantages to multiple-choice computer-based testing compared to paper-based testing; the computer-based multiple-choice format allows test administrators the ability to sample a construct of interest more thoroughly because each question requires less time to answer, allowing the examinee to answer more questions in a given time period (Jodoin, 2003). Another advantage of multiple-choice computer-based testing is that it can be quickly and easily scored and can promptly give student feedback. Nevertheless, computerized multiple-choice testing has been criticized for the belief that multiple-choice items cannot reflect the deeper, higher-level knowledge and skills (Archbald & Newmann, 1988). As a consequence, the current traditional multiple-choice large-scale assessments might not have the capability to assess the more rigorous standards of the Common Core State Standards initiative.

An answer to improving how a students' deeper knowledge and skills can be assessed is using innovative item types (Dolan, Goodman, Strain-Seymour, Adams, & Sethuraman, 2011).

The Development of Innovative Item Types in Educational Assessment

To create new, more comprehensive assessments, the U.S. Department of Education (USED), through the 2010 Race to the Top Competition, awarded grants to two state consortia to develop improved assessments, particularly through the use of innovative item types. The two consortia, The Partnership for Assessment of Readiness for College and Careers (PARCC) and the SMARTER Balanced Assessment Consortium (SBAC), are currently developing more comprehensive online tests based on the Common Core State Standards (CCSS).

Innovative item types allow test item developers to use computer features and functions beyond the ability of multiple-choice item types. For example, innovative item types can include sound, moving media, research resources and tools to interact with these features. The response action can be, for example, clicking to move an item to another part of the screen which allows an answer to be any part of the computer screen opposed to 4 or 5 answer choices. These new innovations target: 1) item format or item structure, 2) response action necessary to perform the task, 3) the media included in the item, 4) item interaction, and 5) scoring method (Parshall et al., 2000).

Most innovative item types can be scored automatically by the computer testing system, allowing the advantages of computer-based testing to remain. Performance tasks and extended constructed response tasks can allow for the capability to assess deeper skills and knowledge (Hambleton, 2000); however, performance tasks require human scoring making scoring more costly and more difficult to score in a standardized way. Additionally, when the number of

students testing is exceptionally large, scoring performance tasks can become an impossibility (Clauser et al., 1997).

Advantages of Innovative Items

There are many advantages to innovative item types compared to multiple-choice item types. The most often cited reason for development of these item types is their capability to measure a deeper knowledge-level and higher-order skills than can be assessed with multiple-choice items (Huff & Sireci, 2001). Additionally, the interactivity the innovative item types allows for a more authentic testing experience for the student and a closer alignment with real-world activities. For example, when the student is able to view a real-world document and answer questions, or construct a representation of a fraction, then the testing experience is closer to the same experience used to learn a skill in the classroom.

When there is a match between the template type and the construct being measured, then measurement is improved. For example, for a line graph item type template the response action requires the examinee to create a line on a graph. This provides an item type that closely measures the construct of graphing a line. Another advantage is that innovative items reduce the possibility that the examinee will guess the correct answer unlike multiple-choice items where guessing is simply done by checking a radio button. Innovative items are more engaging to the students and reportedly students, teachers and policy makers have shown great enthusiasm for these item types and their potential (Strain-Seymour, Way, & Dolan, 2009). Lastly, these items can be automatically scored and provide greater functionality without giving up the advantages of traditional computer-based testing.

With all of the benefits of innovative item types, there is no surprise that these items have been developed and tested in the field on a large scale this year. For example, the Smarter

Balanced Assessment Consortium field tested over 19,000 items and performance tasks during the 2014 field test (Smarter Balanced Assessment Consortium, 2012). The promise of more authentic assessment and providing more comprehensive data to aide in educational instruction and to meet educational objectives has created a large push to create innovative item types for assessment.

Challenges with Innovative Items

Although innovative items show great promise for positive change and innovation in assessment, there have been many challenges with implementation. There are many pieces that must come together to create a smooth and valuable testing experience: The technological computer-capacities of the school and test engine servers, the device used to administer tests (tablet, desktop or laptop), the content and choice of template associated, and psychometric analyses can provide challenges. Within each of these pieces there is cause for some level of concern for test developers and educators.

The cost and time to develop innovative items templates and content is significant for technology developers and item writers. Additionally, the cost and time to build technology infrastructures and bandwidth capacities are a concern among schools and districts. During the 2014 test window, some schools and students experienced slow-down and shut-down issues associated with capacities of computer systems. Also, there was some concern over the additional time it took students to take the test (Gewertz, 2014) since the innovative items were found to take longer to answer than multiple-choice based tests. More scientific evidence is needed in order to weigh the concerns, as well as the benefits, of innovative item types.

There is little empirical evidence to support the idea that innovative items can improve assessment and measurement of knowledge, skills and abilities (Bennett et al., 2010). There is a

plethora of empirical data that supports the value of traditional multiple-choice testing; however, there is little empirical evidence of the value of computer-based, innovative item assessment. To that end, the PARCC Consortia intends to review empirical data in the Fall of 2015 pertaining administration of tests on laptop, desktops and tablets, computer-based testing vs. paper-based testing, among other studies (Partnership for Assessment of Readiness for College and Careers, 2014), illustrating the need and timeliness of this research.

The Use of Innovative Item Types

With the promise of innovative items and their ability to assess the critical thinking skills and higher-order knowledge, it's no wonder there is great interest in establishing empirical evidence regarding the use of innovative items. The Smarter Balance Consortium field tested more than 19,000 assessment items during the 2014 field test to identify which items function well and which items need to be refined in order to assess student performance in a fair and accurate manner (Smarter Balanced Assessment Consortium, 2012). Over one million tests were administered to students in the PARRC 2014 field test and almost 10,000 test questions were being reviewed to determine which items are valid and which will be improved upon based on the 2014 field test (Partnership for Assessment of Readiness for College and Careers, 2014).

The positive reactions from students and teachers to the use of innovative items is an additional reason to ensure these items are comparable by device type. In one study, examinees were asked to provide open-ended statements regarding the use of innovative items. The students used words such as "fun," and "awesome" to describe the experience with these items types. Teachers expressed the relationship between real classroom experiences and the innovative tasks (Strain-Seymour et al., 2009). In another survey of nursing students, it was found that overall the students had positive responses regarding the use of innovative items. The students believed the

items to measure higher-order knowledge and most students found these items to be more authentic than multiple-choice items (Wendt & Harmes, 2009).

Not only have new technologies lead to the development of innovative items, assessments are being administered on various devices (desktops, laptops, tablets). Since teachers are using tablets and other computing devices for instruction, assessments are being administered on those same devices. Educators, from state to teacher level, will want to know that these devices they are using for instruction will have equal comparability when used for administration of assessments.

Computer Device Differences

It is important to note the particular characteristics of the tablet when compared to a computer. The tablet or mobile computer, is smaller than a desktop or laptop, but larger than a cellular phone. Tablets are typically thought of as an iPad or Android tablet. The main characteristics that distinguish tablets from computers are the mobility, the physical size and weight of the device, the size of the screen, touch-screen manipulation, and the function of the keyboard. A tablet can be held for viewing in landscape or portrait orientation or docked in either viewing position (Strain-Seymour, Craft, Davis, & Elbom, 2013). While the lighter weight of the device affords the ability to change the line of sight of the device, consequently viewable screen size is smaller. In a survey of students with frequent access to tablets, students described positive and negative aspects of tablet use, of which included physical and visual discomfort (Sommerich, Ward, Sikdar, Payne, & Herman, 2007).

The touch-screen manipulation also brings with it a measurable difference from the mouse. The mouse stays in one place while the user is out of touch with the mouse, allowing for the user to easily return to the prior position. The touch screen, conversely, demands the user to have

more precision in finger movement to return to the prior location on the touch screen. Additionally, the touch screen eliminates the ability to hover over an icon or object for further information, while the mouse can show a cursor, eraser, or highlighter icon to represent the tool currently in use by the examinee.

Keyboard differences between the tablet and computer are notable. Even though it is possible to connect a tablet to an external keyboard, typically tablet-users type using the tablet touchscreen. With a computer keyboard, the user can feel the keys and edges of the keys and this provides feedback to the user without looking down at the keyboard. The fingers are placed on the "home keys" and experienced typists can find the remaining keys in relation to the home keys without visual reference. Conversely, the touch-screen keyboard necessitates visual feedback and the user must take attention away from the task to ensure touching the correct letter. As a consequence, typing speeds are about 15-30 words per minute (wpm) compared to an average of about 40 wpm on the computer keyboard (Sax, Lau, & Lawrence, 2011).

The Use of Tablets in Instruction and Assessment

As noted previously, PARCC states on their website the intent to review research pertaining to comparability of test administration device next year, illustrating the need to determine the comparability across testing device types (Partnership for Assessment of Readiness for College and Careers, 2014). As districts and teachers determine the hardware needs of their schools, it is important to know the comparability of devices for instruction and assessment. Not only do teachers want to know the value added to their instruction through the use of iPads as a part of daily and weekly instruction, if they make the choice to add iPad (or desktops or laptops) to the classroom, they want to know that assessments administered on the device they choose are comparable, as well.

Since some features of the tablet are different from that of a computer, there are many advantages for using the tablet for instruction. The iPad is manipulated by a touch of a finger on the screen versus using a mouse or touchpad to manipulate the task. As a result of the motor stimulation, students can be more interested in learning, more motivated, and feel more engaged with the construct of interest. For this reason, use of the iPad can keep students interested for longer periods of time during instruction (Agostini, Di Biase, & Loregian, 2010).

Another advantage of the iPad for instruction is the mobility of the device. The iPad screen is 9.7 inches measured diagonally and weighs about 1.5 pounds, depending on the model (Nations, 2014), making it easy to move around the classroom with the device or move the device from room to room. The device can be taken home for use as a reading device or used to collaborate with other students from home or school. Students are able to research and gather information from anywhere at any time.

Additionally, there are thousands of educational applications, better known as “apps,” that can be downloaded for individual use for learning, or can be used by the teacher for collaborative instruction. The Apple App Store states that there are more than 75,000 apps for educational purposes (Apple Online Store, 2014).

It was little wonder that teachers would explore the idea of adding iPads or tablets into their classroom instruction, and many schools have already added iPads to the classroom. New York City Schools have placed over 2,000 iPads in their schools (Hu, 2011) while Los Angeles public schools initiative had begun to place iPads with each student (Blume, 2013). As a consequence of iPads being used for instruction, these same devices are being used to administer large-scale testing. It is important to determine that these devices are comparable to desktops/laptops since the features of the iPad can make for a different testing experience.

In summary, it should be noted that this research was merely a starting point for examining the validity and usability of innovative item types based upon the administration of tests on desktops, laptops and tablets. As a result of this study, further research needs to continue to expand and grow as test creators adjust and improve the performance of innovative item types, and determine if the device utilized to take the test affects performance.

Student motivation and engagement

The relationship between student motivation and assessment is an important matter as large-scale assessment moves toward presumably more authentic and engaging item types and computer device types. Greater engagement with a task and increased student motivation could bring about test scores that will more closely illustrate student understanding of the assessed construct. Understanding of student's experience and perceptions could help to inform decisions of teachers, educators, administrators and policy makers. This research sought to discover how a group of seventh-grade students interacted with different item types and different computer devices by listening to the thought processes as these students interacted with assessment items.

Common to instructional practices is a three-pronged model: the objectives for learning, instructional exercises, and assessment (Susan M Brookhart, 1997). All three influence each other. The instructional exercises are the way to get students to learn the objectives, goals and/or standards set out by teachers, educators or policymakers. Assessment measures student achievement and performance.

The classroom assessment environment is explained as being the social and cultural experience of individuals and how they interpret it. Instructional content, interpersonal relationship and individual thinking and feelings are all a segment of that experience (Susan M Brookhart, 1997). Most assessments, in particular large-scale assessments, are generally

administered to a class or large group of students (Susan M. Brookhart & Durkin, 2003). Nonetheless, individual students have different responses and perceptions about assessments and those differences function within the larger group. Students will perceive assessments and items differently. The differences include how hard the student perceives the task to be, the importance of performing well, and how much effort it will take. Additionally, student interest in the assessment and/or assessment items will effect motivation and consequently, performance on the task(s).

Conclusion

Technological devices are used ubiquitously in and out of the classroom. The use of various technological devices and item types for assessment is a next obvious step. There are purported advantages of using these devices and item types for assessment, including providing schools with flexibility in technology devices they equip students and the motivation and engagement digital devices provide. Although the history and movement of school reform, accountability in schools, as well as the infusion of technology used for instruction and assessment have provided considerable flexibility and choices for instruction and assessment, there is little empirical evidence that various assessment item-types and computer devices are comparable in assessment performance. In today's assessment-driven educational climate, it is important to investigate the comparability of item-types and performance scores across a variety of devices.

Chapter 3: Methods

This chapter described the methods and methodology used in this study in order to compare and explore the use of technology-enhanced item types and various technological devices for instruction and assessment. In this study, the researcher used both qualitative methods and quantitative methods. A cognitive laboratory think-aloud was used to observe student motivation and engagement using both a desktop/laptop and iPad while manipulating and answering multiple-choice and technology-enhanced item types. A two-way analysis of variance (ANOVA) was used for statistical analysis to compare multiple-choice item types administered on a desktop/laptop versus tablet and technology-enhanced item types administered on a desktop/laptop versus tablet.

Quantitative Methods

This research quantitatively addressed the following research questions:

1. Is there a significant difference in seventh-grade student performance on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop?
2. Is there a significant difference in seventh-grade student performance on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?

Quantitative Participants

Participants in this study were seventh-grade students in a Midwestern state taking the annual summative assessment in both English language arts and mathematics. Approximately 450,000 students tested in grades three through eight and high school using the KITE™ computerized assessment system. Approximately 38,000 seventh-grade students took both assessments during the spring of the 2013-2014 school year. Although districts, schools and students could opt out of

taking the assessments this year, in the past taking the assessment has been required and consequently, most districts and teachers opted to have their students take the assessments in order to become familiar with the new testing system.

Quantitative Instrument

The KITE™ test delivery platform was used for the assessments during the prior 2013-2014 school year. The new testing platform was created for one of many reasons, which was to enable technology-enhanced item types to be created, stored, delivered and scored. Since this system has been created to enable test developers to input and deliver technology-enhanced item types, students presumably found the KITE™ platform user-interface similar to the user-interface used in the previous years.

For both ELA and mathematics, one of three test forms were randomly distributed upon examinee login. For each subject, all three forms had the same number of items, sections and parts. Some of the items were the same across each subject-forms, and others were similar items aligned to the same construct knowledge standards. Each section and part required a new password to continue to the next section and part. The test items assessed the state's college-and career-ready standards and the development of the items were guided by the Smarter Balanced (SBAC) Item/Task Specifications.

English Language Arts

The seventh-grade English language arts (ELA) test forms comprised of a total of 55 items. When the student logged in to the test, one of three forms were randomly assigned to that student. There were two sections to each test form with two parts to each section. Part one of each section contained two reading passages with corresponding items. Part two contained "stand-alone" items; meaning the item was not attached to a reading passage. The ELA forms

contained the following types of technology-enhanced item types: Background Graphic, Drop-Down, Matching, Matrix, Multiple Drop Buckets, Ordering, Select Text, and Sticky-Drop Buckets.

The seventh-grade ELA items measured two knowledge constructs, referred to as claims. Table 2, below, describes the constructs measured by each claim for ELA.

Table 2

Claims for the English Language Arts/Literacy Summative Assessment

Claim #1 – Reading "Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts."
Claim #2 – Writing "Students can produce effective and well-grounded writing for a range of purposes and audiences."

(Smarter Balanced Assessment Consortium, 2012a)

Mathematics

The seventh-grade mathematics test forms comprised of a total of 60 items. When the student logged in to the test, one of three forms were randomly assigned to that student. The examinees were given a basic calculator tool for part one of each section. Once they logged into part two of each section, the calculator was no longer available and they were not able to go back to the previous part of the test. The mathematics forms contained the following technology-enhanced item types: Background Graphic, Drop-Down, Matching, Matrix, Ordering, Sticky-Drop Buckets, and Straight Line.

The items measured four knowledge constructs, referred to as claims. Table 3, below, describes the construct measured by each claim for mathematics.

Table 3

Claims for the Mathematics Summative Assessment

Claim #1 – Concepts & Procedures "Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency."
Claim #2 – Problem Solving "Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies."
Claim #3 – Communicating Reasoning "Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others."
Claim #4 – Modeling and Data Analysis "Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems."

(Smarter Balanced Assessment Consortium, 2012b)

Quantitative Group Design

Tests were grouped by content area: English language arts and mathematics and then by form. For the purposes of analysis of innovative item types, items were grouped into two groups: multiple-choice items and innovative items. For purposes of analysis between iPad delivery and PC administration, tests were grouped according to delivery: all Computers (desktop/laptop) will be grouped into one group and iPad in another.

Quantitative Statistical Analysis

A two-way analysis of variance (ANOVA) was conducted to determine the impact of item type (multiple-choice and technology-enhanced) and device type (iPad and desktop/laptop) on student performance. The level of significance was set at $p < 0.01$. If an interaction effect was significant, follow-up tests were conducted to evaluate simple main effects, interaction comparisons, or both. If the interaction effect was not significant, main effects were evaluated. Preliminary analyses was conducted to examine whether the underlying assumptions of the two-way ANOVA were met, and a power analysis for effect size was initially conducted, as well.

Qualitative Methods

Cognitive labs, often referred to as think-aloud protocols (TAP) and cognitive interviewing, have emerged as one of the prominent approaches used to explore the processes the examinees use when responding to assessment items. Typically the method involves observing and collecting information while the examinee is working through an assessment item and talking aloud about their thought process (Beatty & Gordon, 2007). Therefore, it is appropriate to explore the third and fourth research questions qualitatively:

3. Can technology-enhanced item types promote motivation to perform better than multiple-choice item types?
4. Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?

Qualitative Participants

A sample of 10 seventh-grade students were used for this study. The sample was created by requesting volunteers from seventh grade at a small junior high school in a Midwestern state. The researcher attempted to create a gender-stratified group that matched the larger sample.

Qualitative Instrument

Arrangements were made for the researcher to meet each participant at the school the students attend for an approximately one-hour session. The session with each participant took place in a small room at the junior high campus. The Think-Aloud cognitive labs were conducted in a setting closely similar to the classroom testing environment. The room included a laptop computer with an attached mouse and an audio recorder placed upon a desk with two chairs. Distractions were limited as much as possible by clearing the desk of extraneous materials and there were no other people in the room.

The participants were welcomed and "Instructions for Participants" (see Appendix B) were the first page on the test screen. The "TAP Activity: Instructions for Participants" (see Appendix C) were read by the test administrator. The student was informed that the purpose was not to grade the students' performance, but to gather information about the use of technology and how it functions to answer test questions. The participants were given time to ask questions and given the instruction that they could withdraw at any time. Once the student assured the researcher that they understood all instructions, the researcher began the activity and recorded the think-aloud activity using the "TAP Administrator Think-Aloud Observation Form" (see Appendix D) and the "Researcher Recording Form" (see Appendix E). The "TAP Administrator Think-Aloud Observation Form" includes directions for the researcher, a place to record the date/time, student number, and a standardized set of prompts when needed to prod the student to

speak aloud. The TAP Researcher Recording Form includes space to keep a record of observations about examinee statements, posture/position while working with each device, and other notes.

The participants answered 10 assessment items in both mathematics and ELA on both the iPad and a desktop/laptop. The items consisted of three multiple-choice items and a variety of each of the seven technology-enhanced items per subject. The items were a sample of seventh-grade items from the seventh-grade annual assessment pool. Since the testers could have seen the items during their own testing in school, if a student recognized an item, the test administrator had the student move past that item to the next item. A total of four items were skipped by examinees due to recognition from the previous annual assessment. At the end of the session, an information questionnaire (see Appendix E) came up on the computer screen to gather participant's demographic information, computer familiarity and comfort level, and types of technologies used at home and in their classroom.

The researcher requested that the participant talk out loud as they thought through how to answer each question. The researcher refrained from interrupting the participant except to answer participant questions or to remind the participant to think aloud. After the participant finished with the assessments, they had an opportunity to ask any further questions and were thanked for their participation.

Qualitative Data Analysis

The process of data analysis is to organize and categorize information in order to make sense of what has been found. Data analysis involves using three connected processes that include reduction of the data, display of the data and drawing conclusions from the data (Miles,

Huberman, Miles, & Huberman, 1984). Data reduction is the process of transforming and sorting the data so that conclusions can be formed. The display of data organizes the information so that there can be an understanding of what is happening. The drawing conclusions process is what occurs from the data reduction and display so that conclusions can be drawn and verified.

These processes were used throughout the data analysis. The audio recording of the cognitive lab sessions were transcribed and reviewed for accuracy. The transcriptions and the test administrator notes were used for analysis. The transcriptions and notes were categorized first based upon the questions from the assessment and by type of device. This information then further divided by any themes that emerged. As the major concepts and unique characteristics of the data emerged, the data display was manipulated to create tables and graphs to illustrate those themes and find meaning. The conclusions and verification of data occurred throughout this process.

Chapter Four: Data Analysis

A principal purpose of this study was to compare student performance using multiple-choice and technology-enhanced item types administered on a computer and a tablet. An additional purpose was to explore student motivation when testing using the different modes of administration and different item types. Students use a variety of computer devices in and out of the classroom, and these devices are being used not only for instruction but for alternatives to assessment delivery.

Both quantitative and qualitative methods were used to examine the research questions. In order to quantitatively compare student performance on a tablet and computer using both technology-enhanced and multiple-choice item types, a two-way analysis of variance was conducted to determine interaction effects. If an interaction effect was significant, follow-up tests were conducted to evaluate simple main effects, interaction comparisons, or both. If the interaction effect was not significant, main effects were evaluated. To explore student motivation and engagement by item types and device modes, a think-aloud cognitive laboratory was conducted. This approach employs observation and think-aloud protocol, also called "concurrent verbalization" (Ericsson & Simon, 1992).

The following analysis will first describe the findings of the two-way analysis of variance, and then report upon the findings of the think-aloud cognitive laboratory.

Quantitative Results

A two-way analysis of variance was conducted to determine the impact of item type (multiple-choice and technology-enhanced) and computer device type (iPad and desktop) on

student performance scores. The level of significance was set at $p < .01$. The purpose of the two-way analysis of variance was to determine if an interaction exists between the two independent variables on the dependent variable. In this study, the two-way analysis of variance was used to determine if there was an interaction between computer device type and item type on test performance of seventh-grade students, where computer device (iPad and laptop) and item type (multiple-choice and technology-enhanced) were the independent variables, and test performance scores were the dependent variable. The two-way interaction term represents whether the effect of an independent variable on the dependent variable (test performance) is the same for all values of the other independent variable and vice versa (Green & Salkind, 2010).

This quantitative analysis sought to address the first two research questions:

1. Is there a significant difference in seventh-grade student performance on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop?
2. Is there a significant difference in seventh-grade student performance on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?

A 2x2 ANOVA was conducted to evaluate the effects of computer device type and item type on student performance. The means and standard deviations for student performance as a function of the two factors are presented in Table 4 and 5. Participants in this study were seventh-grade students in a Midwestern state taking the annual summative assessment in both English language arts (ELA) and mathematics. The ELA assessments included a total of 55 test items, while the mathematics assessments contained a total of 60 items. For both ELA and math, each examinee was randomly assigned to one of three test forms, an A, B or C form. For analysis, tests were grouped according to subject (ELA or mathematics) and further divided by test form (A, B and C). Each of the forms contains items that assess the same constructs,

although the items on each form may be different. For analysis of item types, items were divided by multiple choice and technology-enhanced and by device type: iPad or laptop computer.

Table 4

ELA
Forms A, B and C
Descriptive Statistics
Dependent Variable: SUM

GRFR	ITEM	DEVICE	Mean	Std. Deviation	N
ELA7A	MC	Computer	30.0197	7.38008	3403
		iPad	28.8216	7.40436	426
		Total	29.8864	7.39142	3829
	TE	Computer	10.8181	2.57357	3403
		iPad	10.3052	2.92829	426
		Total	10.7610	2.61997	3829
	Total	Computer	20.4189	11.07830	6806
		iPad	19.5634	10.83872	852
		Total	20.3237	11.05448	7658
ELA7B	MC	Computer	27.6794	8.17312	3369
		iPad	25.0139	8.37076	431
		Total	27.3771	8.23814	3800
	TE	Computer	9.9234	2.90610	3369
		iPad	9.2181	2.98348	431
		Total	9.8434	2.92316	3800
	Total	Computer	18.8014	10.79109	6738
		iPad	17.1160	10.09401	862
		Total	18.6103	10.72700	7600
ELA7C	MC	Computer	34.8720	8.81675	3337
		iPad	33.5247	9.06025	446
		Total	34.7132	8.85526	3783
	TE	Computer	7.0944	1.94739	3337
		iPad	6.6682	2.14993	446
		Total	7.0441	1.97684	3783
	Total	Computer	20.9832	15.28679	6674
		iPad	20.0964	14.96084	892
		Total	20.8787	15.25043	7566

Table 5

Math Forms A, B and C
Descriptive Statistics
Dependent Variable: Sum

GRFR	ITYPE	Device	Mean	Std. Deviation	N
MATH7A	MC	Computer	19.30	6.814	4792
		iPad	17.70	6.182	710
		Total	19.10	6.757	5502
	TE	Computer	3.52	2.008	4792
		iPad	3.01	1.901	710
		Total	3.46	2.002	5502
	Total	Computer	11.41	9.353	9584
		iPad	10.35	8.654	1420
		Total	11.28	9.272	11004
MATH7B	MC	Computer	18.23	6.867	7321
		iPad	15.69	6.100	605
		Total	18.04	6.845	7926
	TE	Computer	3.13	2.104	7321
		iPad	2.59	2.011	605
		Total	3.09	2.102	7926
	Total	Computer	10.68	9.102	14642
		iPad	9.14	7.971	1210
		Total	10.56	9.030	15852
MATH7C	MC	Computer	18.11	6.059	6846
		iPad	16.44	5.383	975
		Total	17.90	6.004	7821
	TE	Computer	4.47	2.450	6846
		iPad	3.78	2.323	975
		Total	4.38	2.445	7821
	Total	Computer	11.29	8.239	13692
		iPad	10.11	7.569	1950
		Total	11.14	8.167	15642

A preliminary analysis was conducted to determine if the statistical assumptions were met to make sure the two-way analysis of variance is appropriate and yielded a valid result. The first three assumption were met. The first was that one dependent variable was measured at the continuous level. The second was that there were two independent variables and each consist of two or more categorical variables. The third was that there was independence of observations. The fourth assumption considers outliers. Although the data did contain outliers, all outliers within the data were within possible ranges of scores, and since the sample size was so large, the outliers remained within the data set. The assumption of normality was satisfied for all group combinations of student performance scores by item- and device-type, as assessed by visual inspection of Normal Q-Q plots. Although the assumption of homogeneity of variances, assumption that all groups have equal variance was not met, this does not invalidate the use of the F test (Goddard & Lindquist, 1940).

A two-way ANOVA was conducted to test the effect of device- and item-type on student performance. For each of the test forms, the results were as follows:

ELA Form A

There was no statistically significant interaction between the effects of device type and item type on student performance, F (1, 7654) = 2.897, p = .089, partial $\eta^2 = 0$.

Since the interaction effect between computer device type and item type was not statistically significant, an analysis of the main effect for item type was performed receiving a Bonferroni adjustment and being accepted at the $p < .005$ level. There was a statistically significant main effect of item type, F(1, 7654) = 8780.149, p = .000, partial $\eta^2 = .53$. There was a statistically significant main effect of device type, F(1, 7654) = 18.068, p = .000, partial $\eta^2 = .002$.

ELA Form B

There was a statistically significant interaction between the effects of device type and item type on student performance, F (1, 7596) = 19.404, $p < .01$, partial $\eta^2 = .003$.

Since the interaction effect between computer device type and item type was statistically significant, an analysis of the simple main effects for item type was performed receiving a Bonferroni adjustment and being accepted at the $p < .005$ level. There was a statistically significant difference in performance scores for TE and MC item types taken on a computer. For TE and MC item types taken on a computer, mean performance scores were 17.76 points, 99% CI [17.37, 18.14] higher for MC types than TE types, F(1, 7596) = 14037.35, $p = .000$, partial $\eta^2 = .65$. There was a statistically significant difference in performance scores for MC and TE item types taken on an iPad. For TE and MC item types taken on an iPad, mean performance scores were 15.8 points, 99% CI [14.72, 16.88] higher for MC types than TE types, F(1, 7596) = 1421.2, $p = .000$, partial $\eta^2 = .16$. There was a statistically significant difference in performance scores for computer and iPad device types using MC item types. For computer and iPad device types using MC item types, mean performance scores were 2.67 points, 99% CI[1.86, 3.48] higher taken on the computer than on the iPad, F(1, 7596) = 71.76, $p = .000$, partial $\eta^2 = .009$. The simple main effect of device type on mean performance score for TE item types was not statistically significant, F(1, 7596) = 5.024, $p = .025$, partial $\eta^2 = .001$.

ELA Form C

There was no statistically significant interaction between the effects of device type and item type on student performance, F (1, 7562) = 4.064, $p = .044$, partial $\eta^2 = .001$.

Since the interaction effect between computer device type and item type was not statistically significant, an analysis of the main effect for item type was performed receiving a Bonferroni

adjustment and being accepted at the $p < .005$ level. There was a statistically significant main effect of item type, F(1, 7562) = 14296.9, $p = .000$, partial $\eta^2 = .65$. There was a statistically significant main effect of device type, F(1, 7562) = 15.07, $p = .000$, partial $\eta^2 = .002$.

Mathematics Form A

There was a statistically significant interaction between the effects of device type and item type on student performance, F (1, 11000) = 14.833, $p < .01$, partial $\eta^2 = .001$.

Since the interaction effect between computer device type and item type was statistically significant, an analysis of the simple main effects for item type was performed receiving a Bonferroni adjustment and being accepted at the $p < .005$ level. There was a statistically significant difference in performance scores for TE and MC item types taken on a computer. For TE and MC item types taken on a computer, mean performance scores were 15.58 points, 99% CI [15.52, 16.04] higher for MC types than TE types, $F(1, 11000) = 2417.27$, $p = .000$, partial $\eta^2 = .69$. There was a statistically significant difference in performance scores for MC and TE item types taken on an iPad. For TE and MC item types taken on an iPad, mean performance scores were 14.7 points, 99% CI [14.01, 15.7] higher for MC types than TE types, $F(1, 11000) = 3104.54$, $p = .000$, partial $\eta^2 = .22$. There was a statistically significant difference in performance scores for computer and iPad device types using MC item types. For computer and iPad device types using MC item types, mean performance scores were 1.61 points, 99% CI[1.09, 2.12] higher taken on the computer than on the iPad, $F(1, 11000) = 64.72$, $p = .000$, partial $\eta^2 = .006$. The simple main effect of device type on mean performance score for TE item types was not statistically significant, F(1, 11000) = 6.75, $p = .009$, partial $\eta^2 = .001$.

Mathematics Form B

There was a statistically significant interaction between the effects of device type and item type on student performance, F (1, 15848) = 44.321, p = .000, partial η^2 = .003.

Since the interaction effect between computer device type and item type was statistically significant, an analysis of the simple main effects for item type was performed receiving a Bonferroni adjustment and being accepted at the $p < .005$ level. There was a statistically significant difference in performance scores for TE and MC item types taken on a computer. For TE and MC item types taken on a computer, mean performance scores were 15.2 points, 99% CI [14.89, 15.32] higher for MC types than TE types, F(1, 15848) = 32887.83, p = .000, partial η^2 = .68. There was a statistically significant difference in performance scores for MC and TE item types taken on an iPad. For TE and MC item types taken on an iPad, mean performance scores were 13.1 points, 99% CI [12.35, 13.85] higher for MC types than TE types, F(1, 15848) = 2043.55, p = .000, partial η^2 = .11. There was a statistically significant difference in performance scores for computer and iPad device types using MC item types. For computer and iPad device types using MC item types, mean performance scores were 2.55 points, 99% CI[2.0, 3.1] higher taken on the computer than on the iPad, F(1, 15848) = 142.66, p = .000, partial η^2 = .009. The simple main effect of device type on mean performance score for TE item types was not statistically significant, F(1, 15848) = 6.4, p = .01, partial η^2 = .000.

Mathematics Form C

There was a statistically significant interaction between the effects of device type and item type on student performance, F (1, 15638) = 19.6, p = .000, partial η^2 = .001.

Since the interaction effect between computer device type and item type was statistically significant, an analysis of the simple main effects for item type was performed receiving a

Bonferroni adjustment and being accepted at the $p < .005$ level. There was a statistically significant difference in performance scores for TE and MC item types taken on a computer. For TE and MC item types taken on a computer, mean performance scores were 13.64 points, 99% CI [13.44, 13.84] higher for MC types than TE types, $F(1, 15638) = 30563.7$, $p = .000$, partial $\eta^2 = .66$. There was a statistically significant difference in performance scores for MC and TE item types taken on an iPad. For TE and MC item types taken on an iPad, mean performance scores were 12.66 points, 99% CI [12.13, 13.2] higher for MC types than TE types, $F(1, 15638) = 3750.83$, $p = .000$, partial $\eta^2 = .19$. There was a statistically significant difference in performance scores for computer and iPad device types using MC item types. For computer and iPad device types using MC item types, mean performance scores were 1.67 points, 99% CI[1.27, 2.07] higher taken on the computer than on the iPad, $F(1, 15848) = 114.17$ $p = .000$, partial $\eta^2 = .007$. There was a statistically significant difference in performance scores for computer and iPad device types using TE item types. For computer and iPad device types using TE item types, mean performance scores were 0.7 points, 99% CI[.29, 1.09] higher taken on the computer than on the iPad, $F(1, 15848) = 19.57$ $p = .000$, partial $\eta^2 = .001$.

Summary

A two-way analysis of variance was conducted to determine the impact of item type (multiple-choice and technology-enhanced) and computer device type (iPad and desktop) on seventh-grade student performance scores for ELA and mathematics. Each subject contained an A, B and a C form that was randomly assigned to students in the seventh grade. The level of significance was set at $p < .01$. The purpose of the two-way analysis of variance was to determine if an interaction exists between the two independent variables on the dependent variable. In this study, the two-way analysis of variance was used to determine if there was an

interaction between computer device type and item type on test performance of seventh-grade students, where computer device (iPad and laptop) and item type (multiple-choice and technology-enhanced) were the independent variables, and test performance was the dependent variable. For one ELA (B form) and all three mathematics forms, there was a statistically significant interaction between the effects of device type and item type. For these forms, simple main effects for each level of computer device (computer and iPad) was statistically significant, although small, partial η^2 ranging from .001 to .009. Simple main effects across item types (multiple-choice and technology-enhanced) were moderate to large with partial η^2 ranging from .11 to .68. For the two ELA forms (A and C) there was not a statistically significant interaction between the effects of device type and item types on student performance. There was a statistically significant large main effect of item type with partial η^2 of .53 and .65, respectively. There was a statistically significant small main effect of device type with partial η^2 of .002 and .002, respectively.

Qualitative Results

The purpose of this qualitative portion of the study was to explore whether students found using the tablet or computer to be more motivating or engaging and whether students were more motivated and engaged when being assessed using different computer devices and item types. The two research questions used to guide this phase of the investigation were:

3. Can technology-enhanced item types promote motivation to perform better than multiple-choice item types?
4. Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?

A cognitive laboratory think-aloud protocol was used to gather data about the student thought processes (motivation and engagement) while completing the assessment tasks. Observation notes were recorded by the researcher while subjects completed the respective assessments. The researcher sought to understand the thought processes used by students during their experience of navigating the test on both a laptop computer device and an iPad tablet while answering multiple-choice and technology-enhanced item types.

Participants

The cognitive laboratory think-aloud protocol portion of the study was conducted at a small/medium sized junior high in a Midwest suburban city. Ten seventh-grade students volunteered to participate in the study and parent signature forms were collected from each of the ten students. Since all of the students had previously taken the annual summative assessment administered online using the KITE™ software platform all students were familiar with the online user interface and had practiced completing tasks using multiple-choice and technology-enhanced types. Students completed a self-report questionnaire at the end of the cognitive lab. All of the students reported that they had some type of computer device at home and all reported that they used their own home device for homework. All students reported that they liked it when their teacher uses technology to teach (whiteboard, computer, etc.). Student self-reported demographics and information are listed in the table below:

Table 6

Gender	*Ethnicity*	*Student Class Standing*	*Free/Reduced Lunch*
5 Female	9 White/Caucasian	5 Middle	7 Free/reduced
5 Male	1 Hawaiian	5 Top	1 No Free/reduced
			1 did not know

The Process

Students volunteered in response to a classroom announcement by their teacher. Paperwork was sent home to give study information and for parent permission signatures. Appointments were scheduled in cooperation with the principal and teachers at the school according to convenience in student school schedules. Students were administered the assessment at scheduled times throughout the school day in a quiet room with a desk set up with a MAC laptop and an iPad. An audio recorder was placed on the desk. There were no other people in the room and distractions were limited.

When the student arrived for their individual session each participant was welcomed and the researcher checked to ensure that the parent permission form was signed and obtained. Instructions were given and the participant was given a chance to ask questions. Assessments were administered on an iPad and a MAC laptop computer with an attached mouse. Each student took an assessment with 10 items on the iPad and 10 items on the laptop. Each set of 10 consisted of three multiple-choice items and seven technology-enhanced items. These items were selected from the pool of seventh-grade annual summative assessment items. The researcher performed a demonstration on a practice item, and the participant practiced on one item before the cognitive laboratory think-aloud session began. The participants were told to skip items that they recognized from the annual summative assessments. A total of four items were skipped by participants due to item recognition. After the participant finished all items on both computer devices, the student self-report questionnaire appeared after the last assessment item and the participant answered those informational questions out loud while the researcher recorded their answers.

Researcher observational notes and audio recordings were transcribed for each session. Student comments and behaviors were first organized according to student and item on one spreadsheet. On another spreadsheet, student comments and observations were organized by item type and computer device type. Common themes were highlighted in yellow, then common themes across both spreadsheets were highlighted in green. All student self-report questionnaires were summarized on an additional spreadsheet. Although themes were developed from the common responses of participants, individual statements and responses were included in the research as these statements represent critical student perception and cognitive processes. The primary purpose of conducting a think-aloud is to make discoverable the cognitive processes and information processing that happens during a task (Eveland Jr & Dunwoody, 2000; Van Someren, Barnard, & Sandberg, 1994). To make the themes of the group observable, individual statements are reported as examples of specific statements.

Report of Student Statements and Responses

Table 7 below provides a summary of student questionnaire responses to the questions below about using different computer devices and item types while testing:

Which device did you like using the most? Computer / iPad / Both

Which type of item did you like answering the most? Multiple-choice / Technology-Enhanced / /Both

Table 7

Device Type Liked the Most	*Item Type Liked the Most*
5 Liked Computer	7 Liked Multiple-Choice
1 Liked iPad	1 Liked Technology-Enhanced
4 Liked Both	2 Liked Both

The most common reasons for liking one device over another had to do with the ability to use a mouse not only to mark answers, but to guide them in reading the items and answer

choices. Some students remarked that it was difficult to select the radio button with their finger, and that it would help to be able to zoom in order to choose the button. It was observed that most students changed positions with the iPad several times throughout testing. Students would sometimes hold the iPad in both hands for a period of time, then place it on the desk, place it in their lap, and would change holding position when pulling out scratch paper.

The most commonly expressed statements to describe the reason for liking multiple-choice items was because “they’re easier”, took less time and because they could eliminate the other answers and pick the best answer. It was observed that with technology-enhanced item, where possible, the students often used the process of elimination whenever possible. For example, many students held the drop-down icon so they could view all answer choices at the same time and then would choose the best answer. One student commented that they didn’t like to have to “make my own answer” with the technology-enhanced items.

Summary

This chapter provided the results of the two-way analysis of variance and the cognitive laboratory think-aloud activity. The two-way analysis of variance did reveal statistically significant interaction effects for one ELA form and all three math forms. For one ELA (B form) and all three mathematics forms, there was a statistically significant interaction between the effects of device type and item type. Most notable for these forms, was that the simple main effects across item types (multiple-choice and technology-enhanced) were moderate to large for these forms. The two remaining ELA forms had no interaction, but main effects for each. A discussion of possible explanations for these findings is included in Chapter Five.

Additionally, there were positive and negative statements communicated during the cognitive laboratory think-aloud activity. Positive statements about item types and device types included:

flexibility in using different computer devices, variety of item types, and the ability to demonstrate more knowledge. Negative statements included: technology-enhanced items were seemingly harder, difficulty marking answers on the iPad by finger touch, and technology-enhanced item types took significantly more time to answer.

Chapter Five: Findings, Conclusions, and Implications

Introduction

This chapter summarized the research study findings, which were designed to compare student performance and explore student motivation when taking assessments on a computer and a tablet when the assessment consisted of multiple-choice and technology-enhanced item types. Qualitative and quantitative inquiry approaches were used to examine the research questions. A two-way analysis of variance was used to examine student performance scores administered on a tablet or computer using multiple-choice and technology-enhanced item types. A cognitive laboratory think-aloud protocol was employed to explore student motivation and engagement when students were assessed using multiple-choice and technology-enhanced item types on a tablet or computer. The following research questions guided the research study:

Research Questions

1. Is there a significant difference in seventh-grade student performance scores on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop?
2. Is there a significant difference in seventh-grade student performance scores on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?
3. Can technology-enhanced item types promote motivation to perform better than multiple-choice item types?
4. Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?

This chapter is divided into five sections that discuss and interpret the study findings. The introduction section includes a review of the theoretical framework. The second section of this

chapter presents the interpretation of results and implications of the study findings. Section three of this chapter presents the conclusions, section four discusses study limitations, and the last section outlines suggestions for future study. This information should contribute to educator insight as they develop strategies to strengthen the integrity of standardized assessment scores through changes in assessment item types and delivery modes.

Review of Framework

The purpose of this study was to compare student performance and explore motivation and engagement when using multiple-choice and technology-enhanced assessment items when administered on a computer or a tablet. It is argued that these new assessment items will provide better facts for educators and students about the weaknesses and strengths of students (SBAC, 2014). Empirical evidence to support these claims is in the beginning stages and it is critical that the underlying elements of these assertions continue to be verified by research. As instructional and assessment technologies evolve and are added to classroom instruction and assessment, it is important to determine these new technologies are, in fact, adding value to student experience, improving assessment and measurement, and to determine if any differences need to be mitigated.

Interpretation of Results and Implications of Study Findings

The first two research questions addressed the interaction effect between item type and device type. The first question was: Is there a significant difference in seventh-grade student performance scores on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop? The second question was: Is there a significant difference in seventh-grade student performance scores on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?

The two-way analysis of variance revealed statistically significant interaction effects for four of the six forms; one ELA form and all three math forms. For these four forms, there was a statistically significant interaction between the effects of device type and item type. Most notable for these forms, was that the simple main effects across item types (multiple-choice and technology-enhanced) were moderate to large with partial η^2 ranging from .11 to .68, while simple main effects for each level of computer device (computer and iPad) was either not statistically significant or statistically significant, although small with partial η^2 ranging from .001 to .009.

For the remaining two ELA forms (A and C) there was not a statistically significant interaction between the effects of device type and item types on student performance. There was a statistically significant large main effect of item type with partial η^2 of .53 and .65, respectively. There was a statistically significant small main effect of device type with partial η^2 of .002 and .002, respectively.

There are possible explanations that could account for the difference in the interaction effects between content areas of mathematics and ELA. First, the template types for mathematics may be a better match for mathematics item types making the items measure the targeted achievement construct better. Assessments are designed to measure a "domain" of skills, knowledge and abilities and assessments are meant to measure these through a small sample of questions or items (Haladyna et al, 2004). When an assessment item introduces extraneous, uncontrolled variables outside of the domain, or achievement construct, this is called "construct irrelevant variance". Construct irrelevant variance could be a difficulty with new item types until test developers learn which item-type templates best assess particular achievement constructs.

A second plausible explanation for the interaction in all math forms could be that students may have found that ELA item type templates to be more intuitive and had been more practiced as a part of real-world activities. For example, the drop-down item type template is one that is found on many websites. When filling out many types of forms online, drop-downs are used to choose demographic information, and many other types of information. Since this was the first year that students had been tested using technology-enhanced item types, it could be that the students had more prior knowledge of these item types features as compared to math item types. For example, the math item types included graphing items where the student would place a point on a line or graph. These item types would typically not be used in the course of a typical online experience outside of assessments. Additionally, since this was the first year the students were tested using technology-enhanced item types, practice tests for the students to specifically practice manipulating item-type templates were not available to students until the Fall prior to Spring testing. The students may or may not have had extensive practice manipulating and practicing with the various templates. Moreover, comments were made during the cognitive laboratory such as “oh, yeah, I remember how to do this one” and “let’s see, you pull this over here” indicating that prior practice using the templates was beneficial.

One plausible explanation for the statistically significant findings between device types, although small, could be associated with the use of the calculator and the iPad. When answering math items taken on the iPad the students were required to move the calculator to an area where they could still read the question, and/or minimize the calculator and toggle back and forth between calculator and item. It would seem a reasonable conclusion that students who have the use of calculators when answering items that require calculations would have an obvious advantage above students without calculators. However, results presented from the 2000 NAEP

cited in *The Nation's Report Card* (NCES, 2002) revealed an interaction between grade and use of the calculator. There was a correlation with frequent calculator use and lower scores for students in the fourth grade, but for eighth- and twelfth-grade students the reverse was true. Therefore, the use of the calculator appears to be associated with construct irrelevance variance and item type (Haladyna et al., 2004). This indicates that using a calculator is sometimes disadvantageous and it is important to explore this difference with new item types and device types.

The second two questions (Can technology-enhanced item types promote motivation to perform better than multiple-choice item types? Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?) sought to determine whether students found a specific item type or device type more motivating or engaging.

Many noteworthy observations were made to inform the second two questions and are listed below. Usability difficulties issues were observed and are listed next, along with researcher suggested possible solutions.

Observations:

- Most students used the process of elimination to answer questions when possible. Students often used the process of elimination for multiple-choice item types, but also anytime possible the students used the process of elimination for technology-enhanced item types, as well. Some technology-enhanced item types allow for the use of the process of elimination. For example, many students used this strategy with the drop-down item type by clicking on the drop-down icon, viewing each answer choice, and eliminating the answers they thought were incorrect. When students were using the

computer and could use the mouse to hover, the process of elimination was often used with Select Text items by hovering over each option of select text answer choices.

- Students were observed to take more time answering technology-enhanced item types. Although in instances where they were able to use the process of elimination (as with drop-down or select text items), the time was less than other technology-enhanced items but still more than multiple-choice items. Furthermore, many of the students reported that the technology-enhanced items were “hard” when compared to the multiple-choice items.
- Students were very comfortable using both device types and both item types. All students had completed their annual summative assessment using the same assessment user-interface and item-types. When students interacted with the items and devices, there was almost no hesitation or frustration noted or observed. Students expressed that they had previously used the tools of the user-interface (highlighter, calculator, etc.) within the last month on their annual assessments and expressed ease of use.
- Students checked their answers with technology-enhanced item types that allowed for easy checking. For example, once the answer was selected using a drop-down item type, checking the answer was made easy by reading the sentence again with the appropriate answer checked within the text. When the student answered an ordering item type, they could easily read their answers in the order they had placed them in to check their answer.
- There were no internet disconnections or system problems while examinees were testing. Although there have been many concerns about school and district bandwidth and internet reliability, no issues occurred during student cognitive laboratories.

Usability Difficulties and Solutions

The researcher noted usability difficulties during the cognitive laboratory observations and those are outlined below along with suggested possible solutions. Student comments and observations are in the next section of this chapter.

- A minor usability difficulty observed across item types was the inability to select precise spots on the screen when using a tablet. When the radio button on multiple-choice item type was too close to other radio buttons (for example, when the answer choices were only one sentence the radio buttons would be only one line apart), students sometimes accidentally selected the wrong radio button and needed to re-select. In a like manner, radio buttons on the multiple-column item type could be small and too precise to select on the first try. Relatedly, students sometimes had difficulty dragging text or an object to precise spots on the screen. For example, when ordering objects or text with the Ordering item type, students sometimes had difficulty with getting the object or text to "snap" into the intended place on the screen.
- *Solution*: A potential solution is to create larger "spots' on the screen for selecting or dragging and creating a larger separation of selection spots on the screen for tablet test administration. The radio button "select" or "touch" spot on the screen needs to be slightly larger than the actual viewed radio button, allowing for a slight error in the finger touch area on the screen. Similarly, item types that require "snapping" an object or item in place would require a larger area that will allow for the object or item to snap into place. Another solution is to allow for "pinch and zoom" on the tablet so the user can expand the screen area and select a spot on the screen with less precision of finger movement.

- Items sometimes required the student to select an answer more than once in order for the tablet to display the answer on the screen. There seemed to be some delay with the tablet.
- *Solution*: Faster speed with answer choices displaying on the tablet screen after the student has chosen an answer.
- Students often used the mouse on the computer to hover over answer choices (as with the Select Text item type) or used the mouse as a guided reader. This was not an option with the tablet, although students were given access to a tool to use as a guided reader for text passages only.
- *Solution*: Allow for a guided reader tool that would work with all item types.

Conclusion

This study confirms the speed at which technology in instruction and assessment has evolved at a rapid pace. Researchers must continue to keep up the pace in order to ensure that assessments measure the targeted construct and students are not challenged by difficulties navigating item types and device type features. Students were found to be excited about the new technology device choice and new item types. As the evolution of assessment and instructional technology moves forward, the hope is that these device types and item types will bring about a variety of engaging instructional devices, school choice in affordable computer devices, and quality student assessments that target tested constructs.

The first two research questions quantitatively addressed whether there were differences in student performance scores when students were assessed using multiple-choice and technology-enhanced item types on a computer and an iPad tablet. To address these questions, a two-way analysis of variance was performed.

Research Question 1: Is there a significant difference in seventh-grade student performance scores on ELA and mathematics multiple-choice item types when testing on iPad versus desktop/laptop?

For four of six forms (one ELA form and all three mathematics forms) there was a statistically significant interaction between the effects of device type and item type on student performance scores. For these forms, performance scores were higher using MC item types on a computer compared to an iPad tablet. However, the effect was small with partial eta's ranging from .009 to .001. For the remaining two ELA forms, there was no statistically significant interaction between the effects of device type and item type on student performance, however, there was a small statistically significant main effect of device type. No previous studies were found that compared multiple-choice item types taken on a computer/laptop compared to an iPad, however, there are comparable studies that examine screen size and display differences. A study by Bridgman et al. (2003) examined differences in testing on a variety of computerized equipment with differences in screen size, resolution and display. The study participants were college-bound high school juniors that were given a self-report questionnaire and a mathematics and verbal assessment. The students were tested using a variety of screen resolutions, sizes and a delay in the presentation of information (to mimic internet-based delays). A majority of participants in this study reported that scrolling interfered with taking the test. Small score differences on the verbal test were found where passages required scrolling. There were not significant differences in mathematics performance scores. A study by Choi & Tinkler (2002) had similar results. They found students in a K-12 setting to have greater difficulty with reading passages where they had to scroll, and with younger students scrolling interfered more.

These studies, in combination with this current study, intimate that device differences might not be a content issue, but rather difficulty tied to how much the presentation of information can be wholly viewed and to the navigation of the screen and its contents. This leads to the next research question in which the focus is on technology-enhanced item types in which, relatedly, require more complex navigation than the traditional multiple-choice item type.

Research Question 2: Is there a significant difference in seventh-grade student performance scores on ELA and mathematics technology-enhanced item types when testing on iPad versus desktop/laptop?

For four of six forms (one ELA form and all three mathematics forms) there was a statistically significant interaction between the effects of device type and item type on student performance. For three of these four forms, there were also no significant simple main effects of device type on mean performance scores for TE item types. For the fourth of the four forms, there was a statistically significant difference in mean performance scores for computer and iPad device types using TE item types, although it was very small, $F(1, 15848) = 19.57$ $p = .000$, partial $\eta^2 = .001$. For the remaining two ELA forms, there was no statistically significant interaction between the effects of device type and item type on student performance, however, there was a small statistically significant main effect of device type. Although there was very little difference in student performance scores with TE item types across devices, there was a statistically significant difference in performance scores on each device when compared to multiple-choice items item types. Mean performance scores for technology-enhanced item types on both devices, for the most part, were lower than multiple-choice item types. Since TE item types require more navigation and sometimes scrolling, this aligns with the previous studies suggesting that item-

types that require more complex navigation and scrolling to view all the item information might lead to student testing difficulties.

Another factor related to greater difficulty with TE item types is the longer amount of time it takes to answer TE item types (Dolan et al., 2011; Jodoin, 2003). This could account for the lower performance scores when compared to MC item types. In addition to more complex navigation required by the student examinee, in combination with additional test time could lead to greater frustration and less motivation.

Overall, quantitative device-type effects were in general quite small with item-type differences moderate to large. These observations suggest that it would be beneficial to expand our understandings about the elements that influence student performance scores and test behavior. As we continue to design item-type templates and administer tests on a variety of devices, those understandings can make certain that students are responding to assessment content and not being challenged by item-type or device-type difficulties.

The third and fourth research questions qualitatively addressed whether students were more motivated using technology-enhanced item types and if students were more motivated and engaged when the assessment was administered on a desktop/laptop compared to an iPad tablet. To address these questions, a cognitive laboratory think-aloud was performed. The observational and verbal data was collected and used to examine student thinking while interacting with item types and device types during assessment.

Research Question 3: Can technology-enhanced item types promote motivation to perform better than multiple-choice item types?

In general, student feedback was positive with regard to technology-enhanced item types. Students were asked to verbalize their thinking about the item types while testing in Mathematics and ELA. Positive comments were similar to and included:

> "This isn't as boring as most tests. I think more about what I'm doing."
>
> "Oh, okay, I get to make my own line on here. Okay, this is interesting . . ." (Graphing Line item)
>
> "I like these kind because they make it easy to check my answer." (Drop-down item)

Behaviorally, students appeared engaged and interested in answering technology-enhanced items. Students used the technology testing tools such as the highlighter and calculator. Some used scratch paper to calculate answers. Students seemed to be doing their best to answer questions correctly.

Regardless of the positive comments and the illustration of positive behaviors, when each student was asked if they liked multiple-choice or technology-enhanced items more, seven of the ten students responded that they liked multiple-choice item types the best. Students stated that they liked them more because they were easier and took less time to answer. Multiple-choice item types take away the process-of-elimination strategy and require students often times to create their own answer. Although the benefit to test developers is that it enables them to garner more information from student responses these items presumably require higher-order thinking skills, it does require more time and effort on the students' part. Unquestionably, technology-enhanced items were observed to take more time to complete compared to multiple-choice items

in this study and in prior studies (Dolan et al., 2011; Jodoin, 2003). The first three questions of the ten question test for both mathematics and ELA were multiple-choice items. The remaining items were technology-enhanced items. There was a marked difference observed in the amount of time each of the first three items took in comparison to each of the technology-enhanced items. These tests were relatively short at 10 items each, however, in a longer test fatigue could become a factor that introduces construct irrelevant variance. Further examination is needed with regard to the time it takes to test and its' effect on student performance.

Overall, students did express interest in performing technology-enhanced item type tasks, even though students verbally expressed that they liked the multiple-choice item type better. Students were observed to interact and appeared engaged when answering technology-enhanced item types. As student examinees gain more experience with technology-enhanced item types, the amount of time to answer questions should lessen. In addition, test developers will gain greater understanding over time of item types that are best able to measure particular student knowledge, skills and abilities. This will help to shorten test time that includes technology-enhanced item types, however, more examination is needed into how this effects student performance.

Research Question 4: Are students more motivated to engage with item content when testing on a tablet versus a desktop/laptop?

Students' comments about using the iPad while testing were both positive and negative. Positive remarks usually pertained to being able to interact directly with the content, including:

> "I like that I can move this around with my finger and see where my answers are."
>
> "I can set this on my lap, right? I'm comfortable with it on my lap."

Students were observed to easily navigate, choose answers, and move text and graphics on the iPad tablet. Students did not express any frustrations with using the device itself, nor were any observed. Students did express difficulties with regard to screen size with item types where scrolling was necessary in order to view the entire item. For example, when the item included a passage or a graphic that would not fit on the screen some students expressed that they would rather be able to view the entire item at once. Remarks were similar to this statement from one student:

> "It would be really good if the graph and the question could be seen at the same time because I have to keep moving it up and down to see it and the question and that makes it hard."

One student was observed attempting to answer the question from a stem at the top and didn't realize the rest of the question was at the bottom of the page and not viewable without scrolling. This difficulty was not with the tablet only and in some cases, the examinee needed to scroll on the computer/laptop as well. Studies by Bridgeman et al. (2003) and Choi & Tinkler (2002) results indicated that for item types where the full item could not be wholly viewed on the screen and the examinee must scroll, student performance was lower.

When asked which device they like the most, five students stated that they liked the computer/laptop best, one liked the iPad best, and four liked both. When the five students were asked why they like the computer/laptop the most, two stated it was because they liked to use the mouse instead of their finger to navigate. The student that liked the iPad the most stated it was because they used one at home, liked the ease of use, and liked the ability to move it around on the desk and change seating position.

Overall, students seemed positive about testing on the tablet. It is important, however, for test developers to consider the display of test questions and task information when creating item-type templates and items in order to mitigate any negative affect on examinee performance.

In summary, quantitative and qualitative findings suggest that device-type differences are small, and item-type differences medium to large with regard to student performance scores and in student motivation and engagement. However, there are many factors that affect student performance on high-stakes assessment and further understanding and research will help to develop item-types, when combined with various administration devices, that accurately measure targeted constructs.

Study Limitations

First, the quantitative data sample was limited to seventh-grade student performance scores. Consequently, the generalizability of the results to other grades needs to be considered. The results may have been different for younger or older students. Younger students' lack of computer experience may have garnered different results. Conversely, older students may have more computer familiarity yet a greater understanding of the importance of the test, leading to more nervousness.

The low-stakes environment of the cognitive laboratory think-aloud activity could have lead students to be less motivated to answer questions correctly and to fully read and check answers. The testing environment of the cognitive laboratory think-aloud was much more relaxed than the high-stakes environment. Additionally, students were requested to talk out loud as they performed the tasks which is in contrast to the quiet, serious environment of high-stakes testing. It is important to take these differences and limitations into account when interpreting data produced by the cognitive laboratory.

Additionally, the when examining the analysis of variance main effects there is a lack of randomization of the device type and item type, however, the design of the interaction effect within the analysis of variance does mitigate the lack of randomization.

Suggestions for Future Research

The previous sections outlined the results, conclusions and limitations of the study findings. This discussion brings forth areas for future research that could contribute to the quest for testing that accurately measures a targeted construct. This study contributes to that quest, and as innovations in instructional and assessment technologies increases this type of research will be only a beginning.

This research study could be repeated using the future year's student performance measures. Since these student performance scores are based on testing done in the first year of technology-enhanced item types as well as the first year of the use of a variety of device type administration modes, student performance may change as students learn how to better manipulate the items and use the device types. Relatedly, content developers will have the opportunity to study and find the best use of specific item types that best fit content.

Another study could be performed to determine if the lack of feedback that the tablet provides affects student performance. For example, a select text item will highlight a possible answer choice when hovering over the text using a computer, however, there is no hover when testing on a tablet. Relatedly, when the examinee clicks on a tool (for example, the highlighter) when using a mouse or touchpad, the cursor informs the student as to which tool (selector, highlighter, etc.) the student is utilizing through the use of an icon. When the student tries to select an answer with a tool other than selector, the student cannot select an answer. The tablet does not give feedback, whereas the computer with a mouse or touchpad shows the tool icon to inform the student to go back to the cursor.

This particular study did not observe students answering item types that would require extended typing responses. Additional studies are suggested that compare assessments with examinees typing extended response item types on a tablet with a touchscreen compared to a computer with an external keyboard. The touchscreen could be more difficult to type since it gives no feedback when a letter is typed as opposed to the computer keyboard where the keyboard buttons can be felt and pushed. Students would have very little screen space while typing on a tablet since the touchscreen would cover much of the screen. There could be an

additional adverse effect to student performance scores if the examinee were not able to read the question and type a response at the same time.

Another suggested study would consider if holding tablets for a long period of time or setting the tablet on the desk has an effect on student examinee physical comfort. Students often changed positions from holding the tablet on their lap, to setting it on the table, to holding it out while resting their elbows on the table. This could lead to physical discomfort, pain and/or inability to focus on the task, resulting in lower performance scores.

Research into the use of calculators with particular item types, device types and combinations of assessment technologies along with the use of calculators is continually needed (Haladyna et al., 2004). The screen space needed to mitigate construct irrelevant variance is a point of interest, as well as any technology-enhanced item types and the interaction of calculator usage. Calculator usage for specific items and item types may increase or decrease the performance of students (Bridgeman et al., 1995) and further research is needed in this area.

Finally, further study will continue to be needed as assessment and instructional technology continues to evolve. Human and synthetic audio, video and simulation item types will add to the questions, research, and understanding needed to develop assessment items that accurately measure the targeted construct and eliminate item-type and device-type difficulties. As technological innovations continue to change and improve test developers' ability to use a variety of techniques to measure student knowledge, skills and abilities, it is important to examine and research the implications of each new innovation.

References

Atkins, D., Bennett, J., Brown, J., Chopra, A., Dede, C., Fishman, B., & Williams, B. (2010). Transforming American education: Learning powered by technology. *Learning, 114*.

Beatty, P. C., & Gordon, B. W. (2007). The Practice of Cognitive Interviewing. *The Public Opinion Quarterly, 71*(2), 287-311. doi: 10.2307/4500375

Bill & Melinda Gates Foundation. (2012). Innovation in Education: Technology and Effective Teaching in the U.S. Retrieved 5 Aug, 2014, from https://d3e7x39d4i7wbe.cloudfront.net/public/BMGF_Innovation_In_Education.pdf

Birch, C. (2011). Rethinking Education in the Age of Technology: The Digital Revolution and Schooling in America by Allan Collins and Richard Halverson. *American Journal of Education, 117*(3), 433-436. doi: 10.1086/659215

Blume, H. (2013, August 27, 2013). LAUSD launches its drive to equip every student with iPads *Los Angeles Times*. Retrieved from http://www.latimes.com/

Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*(4), 323-340.

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.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10*(2), 161-180.

Brookhart, S. M., & Durkin, D. T. (2003). Classroom Assessment, Student Motivation, and Achievement in High School Social Studies Classes. *Applied Measurement in Education, 16*(1), 27-54.

Center for Educational Testing & Evaluation. (2014). Technology-Enhanced Items. Retrieved 26 May, 2015, from www.ksassessments.org

Chen, J.-L. (2011). The effects of education compatibility and technological expectancy on e-learning acceptance. *Computers & Education, 57*(2), 1501-1511. doi: http://dx.doi.org/10.1016/j.compedu.2011.02.009

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295-320. doi: 10.1191/0265532203lt258oa

Choi, S. W., & Tinkler, T. (2002, April). Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting. In *annual meeting of the National Council on Measurement in Education, New Orleans, LA.*

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*: Teachers College Press.

Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology, 22*(1), 113-120. doi: http://dx.doi.org/10.1037/h0032355

Ericsson, K. A., & Simon, H. A. (1992). *Protocol Analysis : Verbal Reports As Data*. Cambridge, Mass: MIT Press.

Eveland Jr, W. P., & Dunwoody, S. (2000). Examining information processing on the World Wide Web using think aloud protocols. *Media Psychology, 2*(3), 219-244.

Fairbanks, A. M. (2013). Districts Place High Priority on 1-to-1 Computing. *Education Week, 32*, 12-15.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The Effect of Computer-Based Tests on Racial-Ethnic and Gender Groups. *Journal of Educational Measurement, 39*(2), 133-147. doi: 10.1111/j.1745-3984.2002.tb01139.x

Godard, R. H., & Lindquist, E. F. (1940). An empirical study of the effect of heterogeneous within-groups variance upon certain F-tests of significance in analysis of variance. *Psychometrika, 5*(4), 263-274.

Green, S. B., & Salkind, N. J. (2010). *Using SPSS for Windows and Macintosh: Analyzing and understanding data*: Prentice Hall Press.

Grigg, W. S., Daane, M. C., Jin, Y., & Campbell, J. R. (2003). The Nation's Report Card: Reading, 2002.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hart, M. (2015). Maryland District Headed Toward BYOD Programs in Every School. Retrieved 23 June, 2014, from http://thejournal.com

Honeycutt, T. (2013). Technology in the Classroom: The Benefits of Blended Learning. Retrieved 11 June, 2014, from http://www.nms.org/Blog/TabId/58/PostId/188/technology-in-the-classroom-the-benefits-of-blended-learning.aspx

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi: 10.1177/1558689806298224

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 1-15.

MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79-91. doi: 10.1111/j.1467-8535.2005.00524.x

Miles, M. B., Huberman, A. M., Miles, M. B., & Huberman, A. M. (1984). Drawing Valid Meaning from Qualitative Data: Toward a Shared Craft. *Educational Researcher, 13*(5), 20-30.

National Commission on Excellence in Education. (1983). A Nation at Risk: The Imperative for Educational Reform. *The Elementary School Journal, 84*(2), 113-130. doi: 10.2307/1001303

No Child Left Behind (NCLB), 115, Pub. L. No. 107-110, 1425 Stat. (2002).

Odo, D. M. (2012). Computer Familiarity and Test Performance on a Computer-based Cloze ESL Reading Assessment. Computer Familiarity and Test performance on a Computer-based Cloze ESL Reading Assessment(3), 18-35.

Parsad, B., Jones, J., & Greene, B. (2005). Internet Access in US Public Schools and Classrooms: 1994? 2003. ED TAB. NCES 2005-015. *US Department of Education*.

Parshall, C. G., & Harmes, J. C. (2007). *Designing templates based on a taxonomy of innovative items.* Paper presented at the Proceedings of the GMAC Conference on Computerized Adaptive Testing. Online-Dokument: http://www.psych.umn.edu/psylabs/catcentral/(29.9.2010)

Parshall, C. G., & Harmes, J. C. (2009). Improving the Quality of Innovative Item Types: Four Tasks for Design and Development. *Journal of Applied Testing Technology, 10*(1).

Partnership for Assessment of Readiness for College and Careers. (2014). About PARCC. Retrieved 25 July, 2014, from http://www.parcconline.org/about-parcc

Pressey, B. (2013). Comparative analysis of national teacher surveys. New York: The Joan Ganz Cooney Center at Sesame Workshop.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series. NCES 2005-457. *National Center for Education Statistics*.

Sax, C., Lau, H., & Lawrence, E. (2011). *LiquidKeyboard: An ergonomic, adaptive QWERTY keyboard for touchscreens and surfaces.* Paper presented at the ICDS 2011, The Fifth International Conference on Digital Society.

SBAC. (2014). Parents & Students. Retrieved 20 May, 2015, from www.smarterbalanced.org

Smarter Balanced Assessment Consortium. Smarter Balanced Assessment Consortium. Retrieved June 15, 2014, from http://www.smarterbalanced.org/about/

Smarter Balanced Assessment Consortium. Technology. Retrieved 3 Aug, 2014, from http://www.smarterbalanced.org/smarter-balanced-assessments/technology/

Smarter Balanced Assessment Consortium. (2012a). Claims for the English Language Arts/Literacy Assessment. Retrieved April 20, 2014, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Smarter-Balanced-ELA-Literacy-Claims.pdf

Smarter Balanced Assessment Consortium. (2012b). Claims for the Mathematics Summative Assessment. Retrieved April 20, 2014, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Smarter-Balanced-Mathematics-Claims.pdf

SETDA. (2015). Priorities: Online Assessment. Retrieved 26 May, 2015, from www.setda.org

Sommerich, C., Ward, R., Sikdar, K., Payne, J., & Herman, L. (2007). A survey of high school students with ubiquitous access to tablet PCs. *Ergonomics, 50*(5), 706-727.

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.

Taborn, W. (2011). Tower School's 1:1 Program Brings the iPad 2 to the Elementary Classroom. Retrieved 2014, from http://thejournal.com

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer-based TOEFL.

U.S Department of Education. Elementary and Secondary Education Act. Retrieved May 24, 2015, from http://www.ed.gov/esea

United States Department of Commerce. (2011). Fact Sheet: Digital Literacy. Retrieved 11 June, 2014, from http://www.commerce.gov/news/fact-sheets/2011/05/13/fact-sheet-digital-literacy

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical guide to modelling cognitive processes* (Vol. 2): Academic Press London.

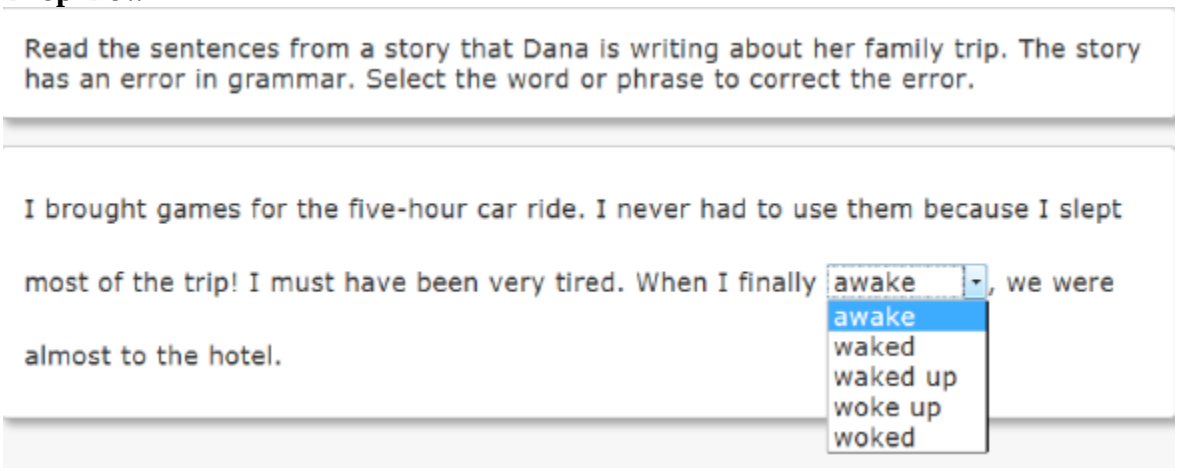
Vu, P., McIntyre, J., & Cepero, J. (2014). Teachers' Use of the iPad in Classrooms and Their Attitudes toward Using It.

Zenisky, A. L., & Sireci, S. G. (2002). Technological Innovations in Large-Scale Assessment. *Applied Measurement in Education, 15*(4), 337-362. doi: 10.1207/S15324818AME1504_02

APPENDIX A

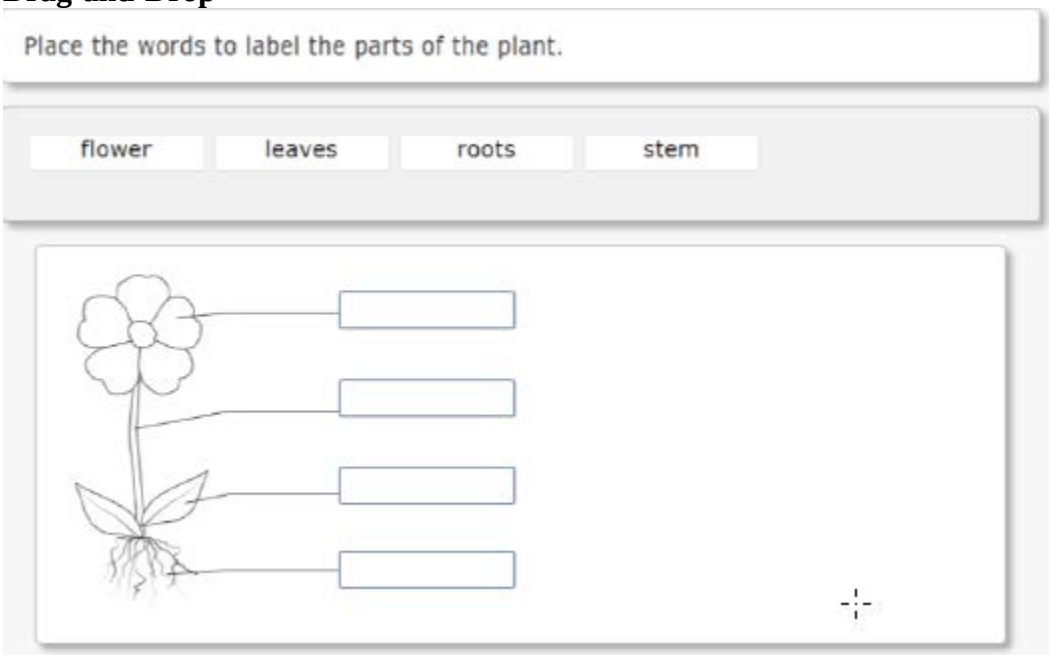
Examples of Technology-enhanced Item Types (Center for Educational Testing & Evaluation, 2014).

Drop-Down



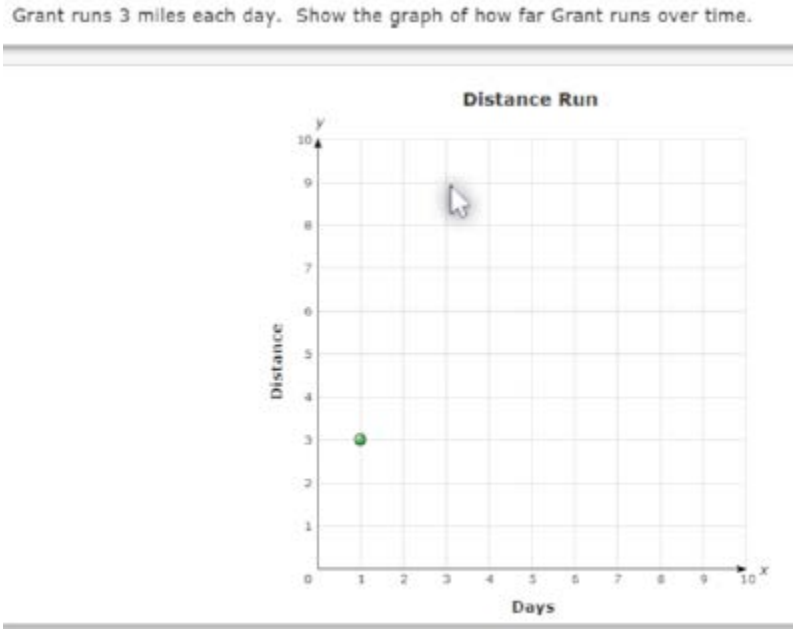
Action: Examinee clicks to select the appropriate answer from the drop-down menu.

Drag-and-Drop



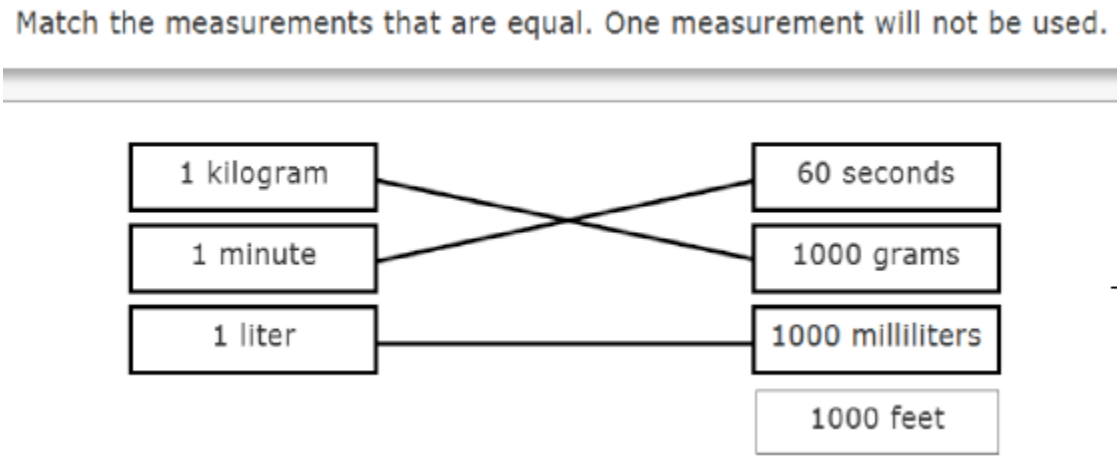
Action: Examinee drags the labels to the appropriate answer area.

Graphing



Action: Examinee clicks to create one point. A second click will create a line between the two points.

Matching Lines

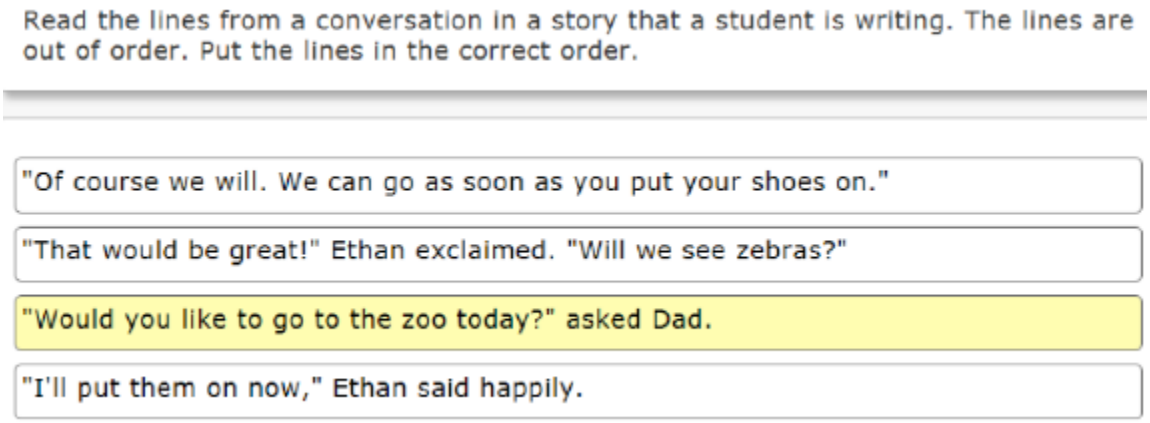


Action: Examinee clicks to choose an option on the right, then clicks to select the matching option on the right to form the matching line.

Multiple Columns

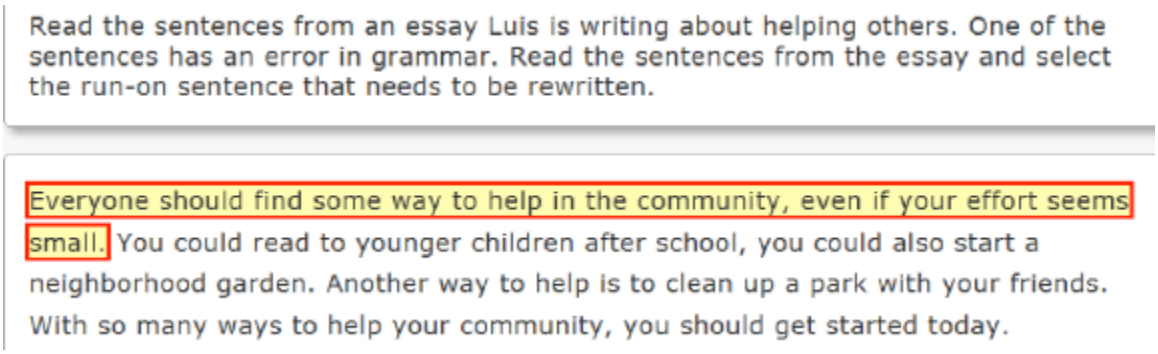
Action: Examinee clicks to select any combination of responses.

Ordering



Action: Examinee clicks to select a response option and then drags it to the appropriate ordering. Ordering can also be from left-to-right.

Select Text



Action: Examinee clicks to select a response by clicking anywhere within the given test space.

APPENDIX B

Instructions for Participants

You are being asked to complete assessment items on two different computer devices – a computer and a tablet. As you complete each item, you will be asked to "think" aloud. In other words, you will say your thoughts aloud as you complete the items. As you talk, your comments will be recorded.

The purpose of this activity is to get more information about what you are thinking while you complete the items. This information will help us improve the tasks and items for other students.

Do the best job you can on the task. Remember, though, that <u>you will not be graded</u> and whether or not you get the answer correct will not be recorded in any way. This is just for information about the features of the tasks.

Before we begin, I will demonstrate using a practice item, and then you will practice on one item.

Also remember, if at any time you feel uncomfortable you are free to stop the task.

APPENDIX C

TAP Activity: Instructions for Participants

To be read by the test administrator:

"You are being asked to complete assessment items on two different computer devices – a computer and a tablet. As you complete the each item, you will be asked to "think" aloud. In other words, you will say your thoughts aloud as you complete the items. As you talk, your comments will be recorded".

"The purpose of this activity is to get more information about what you are thinking while you complete the items. This information will help us improve the tasks and items for other students".

"Do the best job you can on the task. Remember, though, that you will not be graded and whether or not you get the answer correct will not be recorded in any way. This is just for information about the features of the tasks".

"Before we begin, I will demonstrate using a practice item, and then you will practice on one item".

Perform the practice items here.

"Okay, good. Let's go ahead and begin. Remember to speak out loud about what you are thinking as you answer the items".

"Also remember, if at any time you feel uncomfortable you are free to stop the task. Do you have any questions?"

APPENDIX D

TAP Administrator Think-Aloud Observation Form

Observer's Initials: _____Device/Subject: _____

Date/Time: _____Student ID: _____

You may prompt the student to think aloud my making statements, such as:
"What are you thinking?" or "Please think out loud."

If the student asks a question make neutral statements, such as:
"What would you do if you were in your regular classroom and this was a real test?" or "What do you think?"

Read the Instructions for Students. Did the student have questions about the instructions? Record here:
_____

Observe the student as each component of the task is completed. Record your observations about:

- **Task**. Describe any difficulty the student had completing any part(s) of the task.
- Did the administrator have to cue students to keep talking? Why? At what point during which task?
- How many times did you, the researcher, cue the student to keep talking? What did you say to keep the student talking?
- **Student's Questions/Comments**. Record any question/comment the student had for the administrator and administrator response.
- **Difficulties**. Where, when and why did the student have any difficulties?

APPENDIX E

Researcher Recording Form

Circle one: iPad Computer
Participant Identification:

Question 1:	Participant Think-Aloud notes:
Question 2:	Participant Think-Aloud notes:
Question 3:	Participant Think-Aloud notes:
Question 4:	Participant Think-Aloud notes:

Observation recording rows continue through all items.

APPENDIX F

Student Questionnaire
(To be filled out by student by computer after Think-Aloud Activity)

Date: _____ **Time**: _____

Gender: M / F **Ethnicity**: White / African-American / Hispanic / Other

Describe your student standing in your class: Top / Middle / Bottom

Do you receive free or reduced lunch at school? Y / N

Do you have a computer at home? Y / N

If yes, what types of computer devices do you use at home?
If yes, approximately how many hours do you spend using your home computer device?
Do you use this device(s) for home work?

Do you have a computer device at school? Y / N

If yes, what type(s) of device? _____

If yes, do you have a device to use just for your own purposes or do you share?
_____

Do you like using a computer device for your homework? Y / N

Do you like it when your teacher uses technology (iPads, Computer, Whiteboard, etc.) in the classroom? Y / N

Does your teacher use an iPad or tablet for instruction? Y / N

Which device did you like using the most? Computer / iPad / Both

Which type of item did you like answering the most?
Multiple-choice / Technology-enhanced / Both