A COMPARISON OF COMPUTER BASED TESTING AND PAPER AND PENCIL TESTING IN MATHEMATICS ASSESSMENT

Tara McClelland
University of North Georgia
tmac0527@yahoo.com

Josh Cuevas
University of North Georgia
Author Note: Correspondence concerning this article should be addressed to Tara McClelland at:
tmac0527@yahoo.com

Abstract
Today's schools turn to computers for all aspects of learning, including assessment. While advantages to computer testing do exist, the comparability between paper pencil tests (PPT) and computer-based tests (CBT) must be considered. This study examined whether the testing medium impacts student performance in math assessment by addressing three questions. First, does a test mode effect exist, as evidenced by mean score difference between a CBT and a PPT? Second, does question type: multiple choice, constructed response, or extended response, relate to student performance? Third, does either gender or computer experience and familiarity impact CBT and PPT scores? Eighty 6th grade students took math tests with half of the questions on a PPT and half of the questions on a CBT. A computer familiarity survey was completed prior to the unit tests. Significant differences were found for one of the unit tests and for some of the question types.
Keywords: paper-pencil tests, computer-based tests, test mode effect, mathematics assessment, item construct

Introduction

Technology plays an increasingly integral role in education today. This is evidenced in the daily instruction from teachers, to student interactions with the curriculum, and even to homework completion. One component of this increasingly technology-driven classroom is CBTs, computer based tests. The trend of CBT in lieu of the traditional paper and pencil tests, PPT, is witnessed across all academic subject areas. It is a trend also experienced at both the individual classroom level and at the state level on standardized tests.

Advantages of Computer Based Testing

The use of CBTs has numerous benefits for both students, teachers, and educational systems. One such benefit is the real-time scoring and immediate feedback provided by CBTs (Jeong, 2012). With reduced grading time, teachers are able to increase their teaching time (Eid, 2005). The individualized student data generated from CBTs can facilitate teacher instruction to be more strategically directed to enhance individual student goals (Johnson & Green, 2006). Due to the ease at which they can be manipulated, numerous test versions can easily be created, thereby increasing test security (Bodmann & Robinson, 2004; Poggio, Glasnapp, Yang, & Poggio, 2005). Additionally, the ease with which these tests can be manipulated lends itself to increased student control over testing and a medium that is easier to individualize for testing accommodations for students with learning disabilities (Bodmann & Robinson, 2004; Flowers, Kim, Lewis, & Davis, 2011; Jeong, 2012). Finally, the move to CBTs provides a more cost-effective way to assess students, reducing paper costs, administration costs, and scoring costs (Jeong, 2012; Threlfall, Pool, Homer & Swinnerton, 2007).

The Importance of Comparability

While the advantages of CBTs are numerous, a foremost consideration must be whether or not CBT and PPT are equivalent in their assessment of content knowledge and understanding. It must be assured that they reflect a student's content proficiency, not computer proficiency (Puhan, Boughton, & Kim, 2007). As previously noted, the administration of computerized tests is less expensive than the administration of paper and pencil tests, but this is only true after the school or district is in possession of the devices. And while technology has become commonplace in schools, most schools and districts have not made a complete switch to computerized assessments due to financial or infrastructure constraints, thus CBTs and PPTs co-exist during this transitional stage and must be comparable in their assessment of student performance (Minnesota Department of Education, 2012). This dual testing mode is especially significant when used in large scale programs, such as statewide assessments (Kim & Huynh, 2007). Results from statewide assessments are aggregated across individuals taking tests using different modes, so scores from the two testing modes must be interchangeable (Bennett et al., 2008; Poggio et al., 2005). Additionally, as CBTs are increasingly used in other high stakes testing situations, such as certification testing, licensing, and graduate exit exams, comparability is paramount (Keng, McClarty, & Davis, 2008; Puhan et al., 2007). Finally, comparability is critical for tests, such as the PISA, the Programme for

International Student Assessment, which has transitioned to CBT, as it is used for longitudinal data (Logan, 2015).

Previous Comparability Studies

Previous studies have been conducted to determine comparability between the test modalities and have resulted in varied findings; some indicate that CBT and PPT scores are comparable, while others indicate a performance advantage for either CBT or PPT. Bayazit and Askar (2012) examined a test administered to Turkish university students across the two testing modes. While the mean score for the PPT was slightly higher, it was not a statistically significant difference. Similarly, another university study in which students were alternately administered two tests over the course of two weeks in CBT and PPT versions found no testing modality difference for test scores (Bodmann & Robinson, 2004). While both of these were small scale university studies, similar results have been found in large scale testing conditions.

An examination of a PRAXIS test, measuring reading, writing, and math, found only a small effect size (< .20) and determined overall test mode comparability (Puhan et al., 2007). Deeper examination of the test explored individual item level comparability. Differential Item Functioning (DIF) analysis was used to determine if the item measured different abilities for different groups. Despite the score differences between the two testing mediums on the math and reading portions of the test, the test items were found to be comparable. A mode DIF on items in the writing portion was found to exist. Kim and Huynh (2007) found no statistically significant score difference for students taking End of Course (EOC) exams for Biology and Algebra. Though the Algebra mean scale score was higher for PPT and ANOVA revealed a significant mode effect, the effect size was determined to be 0.17 and thus small. Poggio et al. (2005) examined 7th grade math scores from a statewide assessment in Kansas in which 48 schools administered parallel CBTs and PPTs. No statistically significant score differences were found between the testing modes. Though 9 of the 204 items were found to function differently between the mediums, it did not impact composite scores. Likewise, Johnson and Green's (2006) analysis of middle school students' performance on a mathematics test from the National Curriculum for England garnered similar results. No statistically significant differences existed on overall performance, indicating CBT and PPT comparability. A similar finding was observed in an examination of student performance on a test modeled on the end of course assessment used in England (Threlfal et al., 2007). While these studies demonstrate comparability between testing modes, other studies have found higher scores on one mode over the other.

Eid's (2005) analysis of 5th grade females completing a math problem solving test on alternately administered CBTs and PPTs showed a statistically significant difference occurring between the two testing modes, favoring the CBT. Another study examined the mode effect of CBT and PPT on a multiple-choice test taken by students with ADHD (Lee, Osborne, & Carpenter, 2010). A statistically significant difference was found favoring those students in the CBT condition. Likewise, Clariana and Wallace (2002) examined the posttest scores of students in a CBT or PPT condition on a 100-question multiple choice test. The results showed a statistically significant difference with a large effect size in favor of the computer test takers. These studies show an advantage for CBT, but numerous other studies reflect one for PPT.

Flowers et al. (2011) examined data for students in grades 3-8 with a read aloud accommodation from a large-scale statewide assessment of reading, math, and science. Effect sizes ranged from small to large, but favored the PPT condition in almost all grades and subject areas, even though the CBT condition afforded students greater control over their read aloud accommodations. Another study considered the test performance for students with a learning disability taking the Mod-MSA, an alternative to the Maryland School Assessment. Students were assigned to alternatively matched groups based on their previous year's MSA scores, and placed in a computer based testing condition or a paper testing condition. Students taking the PPT scored higher than average on the PPT, with the main effects significant for both reading and math (Taherbhai, Seo, & Bowman, 2012).

Studies involving students in the general education population have also had results favoring test performance on a PPT. A study of 6th graders' scores in four academic areas found that students scored higher on the PPT for all four academic areas, with significant differences in two content areas, Korean language arts and science (Jeong, 2012). In a second study, the scores of 804 6th graders were analyzed on a mathematics-only CBT and PPT (Logan, 2015). The results determined a statistically significant difference for half of the domain areas, with PPT scores higher than CBT scores.

Such favorability with PPTs has also been seen on statewide examinations. Keng et al. (2008) looked exclusively at item level differences on a statewide assessment in Texas. Results determined significant differences for items in all subjects for 8th grade and for math and reading in 11th grade. These differences generally were in favor of PPTs. An analysis of the Mathematics Minnesota Comprehensive Assessment Series III, or MCA-III revealed

higher student performance on the PPT for all grades (Minnesota Department of Education, 2012). Though effect sizes were small, score differences were statistically significant for grades 3-7. Finally, results from the Math Online study, or MOL, were considered (Bennett et al., 2008). The mean scale score for the two modes was statistically significant in favor of the PPT version, though the effect was moderated as computer familiarity increased, as measured by a background questionnaire. The computer-based scores reflected greater variety than the paper based tests.

Explanations for Score Differences

Explanations for score differences have been proffered and examined in numerous studies. Issues that can impact student performance and test score have included testing duration, content domains and problem-solving strategies, testing administrative factors, gender, and computer familiarity.

Duration. Bayazit and Askar (2012) found no difference in student performance on the two testing modalities despite a statistically significant difference in duration. The average time spent on the online test was 40.53 minutes, while the paper pencil test was 34.26 minutes. Another study found increased performance on PPTs along with increased duration on the PPT with the average time of the PPT at 17.49 minutes and the CBT at 15.16 minutes (Dimcock, 1991). Similarly, Bodmann and Robinson (2004) found a statistically significant difference in test duration with the PPT taking an average of almost 4 minutes longer than the CBT. In this case though, the dual test modes resulted in comparable student performance. In an effort to further examine the test mode effect on test times, a second experiment was conducted in which three different CBT conditions were created, all allowing varying levels of flexibility for review and answer changing. In the CBT condition that most closely resembled the PPT condition with the greatest level of flexibility test duration did increase but still did not affect test scores. Studies that have examined test taking duration have resulted in mixed results, with some revealing increased duration for CBT and others revealing increased duration for PPT.

Domain and strategy. Domain effect was examined in the analysis of a mathematics test that was broken into five domains (Logan, 2015). Means scores were higher on PPT versions, and significant differences were found in three domains: Whole Number, Algebra Patterns, and Data and Probability, favoring PPT takers. In all of these domains, the PPT allowed affordances not available to those using CBT. These include students' ability for working out multi-step representations of their work, such as drawing models. It is suggested that the multi-step nature of the problems could cause an increase in cognitive load when answered on the CBT. Such findings were similar to another study that found student performance on PPTs to be higher than CBTs, and additionally found mathematical domains that benefited from the affordances given to PPT users (Keng et al., 2008). These domains included Linear Functions, Geometric Relationships, and Spatial Reasoning. The added burden of transferring drawings from screen to paper increased the difficulty of the CBT. The ability to draw on items on the PPT was an additional support given to PPT users.

Problem solving strategies not just tied to specific domains are also seen to be impacted by the testing medium. Johnson and Green (2006) analyzed the test results from the mathematics test of the National Curriculum for England and found numerous instances of strategy influenced by test mode, such as the ability of students to rotate angles on paper but not on the computer, the prevalence of employing portioning strategies to solve computation problems on paper, and the affordance of creating models to solve problems. Strategies also moderate according to question construct, such as constructed response (Bennett et al., 2008). When answering constructed response on a CBT, students have a more difficult time answering in an alternative way, such as providing a diagram, a beneficial option that would be available on a PPT. Another affordance granted PPT users is the ability to see question and problem relationships within the test (Johnson & Green, 2006). During testing administration, it was observed that PPT users had a greater tendency to preview the entire test, enabling them to note question relationships. It was also observed that PPT users were more likely to review and amend answers. It has been suggested that CBTs permit allowances for math problems in which elements must be arranged to determine a solution, while PPTs permit allowances for math problems in which geometric shapes must be drawn and measured (Threlfall et al., 2007). The fact that affordances are given based on the testing mode calls into question whether either test mode is assessing what was meant to be assessed.

Testing administrative factors. Testing administration factors have also been considered for their impact on test mode effect. These factors include testing processes unique to the test mode that potentially impact the examinee's experience. Bayazit and Askar (2012) found that while CBT and PPT scores were comparable, students took longer to complete the CBT due to the physical dimensions of the computer. The amount of information shown on screen is a fraction of what is shown on paper, causing students to use their time to scroll to see what would be completely visible on paper. Other studies have found scrolling to be of concern as well. The lower resolution of CBTs makes it more difficult to read long passages; however, if resolution is adjusted to

be higher, the result is increased scrolling, again making the reading of long passages more difficult (Jeong, 2012).

Another study exclusively explored CBTs under three conditions, in regard to screen size and resolution (Bridgeman, Lennon, and Jackenthal, 2003). It was observed that screen size and resolution impact the amount of text visible to the reader which impacts the amount of scrolling required. The increased scrolling did not impact math scores, but it did negatively impact the verbal scores. Likewise, a study found comparable overall CBT and PPT scores with item level and domain level differences favoring PPT (Keng et al., 2008). Sizable graphics on the math test resulted in all the answer choices on the CBT being unable to fit on the same screen, resulting in students needing to go to the next page to see the remaining answer choices. In comparison, the PPT users were able to view the large graphic and all answer choices on one page. Additionally, reading test scores were found to be comparable on the CBT and PPT versions, yet favored the PPT version on item level and objective level questions when scrolling was required or reading passages appeared on split screens. Poggio et al. (2005) found comparable CBT and PPT scores in his study of a statewide math assessment, but also found 9 of 204 items that functioned differently. These items were determined to be more difficult in CBT mode, as they contained large items that required scrolling.

Gender and familiarity. Gender and computer familiarity have been considered to account for score differences. Jeong (2012) found the male mean CBT scores were significantly different in comparison to PPT scores in only one of four academic areas whereas, females had statistically significant lower scores on three of the four academic areas for CBT scores, suggesting a gender gap on computer usage. Students in this study had been receiving weekly computer literacy lessons for five years; however, the students were all new to CBTs. Bennett et al. (2008) conversely found gender did not account for test mode differences on the lower scoring CBT, but that increased computer familiarity moderated the test mode effect for CBTs. While Clariana and Wallace (2002) found statistically significant score differences favoring CBT, they also concluded that gender was not a factor but instead, computer familiarity was the most fundamental issue in score differences, with previously identified higher abled students performing even higher on the CBT. Additionally, in terms of gender, no differences were found in other studies (Lee et al., 2010; Poggio et al., 2005). Analyzing the scores of middle school students from a national mathematics test in England, boys were found to be more likely to fail to submit an answer when testing on either mode (Johnson & Green, 2006). This difference was moderated when testing on the computer.

Student Perception and Preference

Student perception and preference between the testing modalities has been examined. Online tests have been described as providing a more relaxed testing environment and a more enjoyable visual medium (Bayazit & Askar, 2012). On the other hand, CBT examinees also noted that it was more difficult to answer long questions because of the typing and they found they lost motivation and time due to screen disturbances and noises. Johnson and Green (2006) observed a similar finding regarding a relaxed environment. They suggested that answering test questions on a computer provides a less personal experience. Once the child submits an answer on the computer, they no longer see that response, as they do with a paper test. On a paper version of a test, a child continues to be exposed to their previous questions, including those they may have struggled with, leading to a more stressful environment. In contrast, the computer provides an "out of sight, out of mind" experience, leading to a less stressful testing environment. This notion was reinforced by the majority of students who felt that computer based questions were easier than paper based questions (Bayazit & Askar, 2012). Both of these studies resulted in comparable CBT and PPT scores (Bayazit & Askar, 2012; Johnson & Green, 2006).

Other studies have found that preference and perception did not correlate to performance. In the analysis of students with read aloud accommodations, it was found that students preferred the CBT over the PPT and predicted that they performed better on the CBT, despite a higher overall PPT score (Flowers et al., 2011). Conversely, the examination of CBT and PPT scores by students with AD/HD found a statistically significant difference in scores, favoring CBT examinees, even though the majority of test takers expressed a preference for paper and pencil for its affordance of being able to write on the test (Lee et al., 2010). In a study of 6th grade students randomly assigned to an online or paper number line estimation task, an analysis of preference and performance concluded that preference did not result in increased performance. It was determined that even though 71.8% of the students preferred the tablet to the paper version of the task, the performance on the two modes was comparable (Piatt, Coret, Choi, Volden, & Bisanz, 2016).

Research Questions

With the ever-increasing presence of computers in society, schools are turning to computers for all aspects of learning, including assessment. Numerous studies have explored the equivalency of computer based testing and

paper and pencil testing, CBT and PPT, respectively. The purpose of this study will be to further examine whether the testing format impacts student performance, specifically in math assessment. The initial question will be whether there is a test mode effect, as evidenced by a difference in mean scores between a CBT and a PPT.

While the mean scores can provide an overall understanding of CBT and PPT equivalence, an examination of the impact of test question type will need to be considered. Increased focus on students' ability to not only correctly perform math equations, but to interpret and solve problems, as well as justify the responses has led to the addition of constructed response items and extended response items on math tests. Therefore, the second research question will seek to determine if the question type, specifically, multiple choice, constructed response, and extended response, has an effect on student performance on CBT and PPT scores.

When score differences between CBT and PPT have occurred, previous research has sought to provide explanations. While item level differences, such as question type, may provide one explanation, additional factors may contribute to score differences. These factors include gender and computer experience. The final research questions will examine if gender has an impact on CBT scores and if computer experience and familiarity is related to CBT scores.

Method

Contextual Factors

This study was conducted in a school district in north Georgia. It is a large district in a community that is experiencing growth. Specifically, it is the 7th largest district out of the 180 districts in Georgia and is part of a community that is the 11th fastest growing community in the United States. The county, with a population of 221,009, has a median household income of $88, 816 and a median housing value of $267,300 (US Census Bureau, 2016). The school district consists of 37 schools, including 21 elementary schools, 10 middle schools, and 7 high schools. The total student enrollment is 44,673 and is comprised of the following: 15.21% Asian, 3.39% Black, 12.94% Hispanic, and 65.22% White students. The district has been experiencing increases in diversity, with the largest increase in Asian students. Free or reduced lunch is received by 17.65% of the students in the district.

The school at which the study was conducted is more racially and ethnically diverse in comparison with the district, with 25% Asian, 8% Black, 26% Hispanic, and 26% White students. The school has a lower socio-economic profile than the district, with 30% of the students receiving free or reduced lunch.

Participants

The participants in the study were 80 6th grade students, 34 boys and 46 girls, from four math classes, comprised of two co-taught classes, one on-level class, and one advanced class. One co-taught class had 20 students, and the other one had 17 students. The on-level class had 18 students, and the advanced class had 26 students. There were 12 special education students and 4 gifted and talented students in the classes. The 6th grade math curriculum does not have a singular focus, such as algebra or geometry, but instead draws on the following five mathematical domains: ratios and proportional relationships, the number system, expressions and equations, geometry, and statistics and probability. The demographics of the students in the classroom were similar to those of the school.

Materials/Measures

Testing materials. Two unit tests were utilized for the study. The questions on the tests came from a bank of test questions that were created by teachers in the county. All 6th grade students in the county are administered tests that are created from the test bank of questions. The tests are required to contain multiple choice, constructed response, and extended response items. The first test, Unit 5 Test, assessed student knowledge, understanding, and application of one step equations. The second test, Unit 6 Test, assessed student knowledge, understanding, and application of area and volume. Each of the two tests were taken in the students' regular math classroom or in the co-teacher's classroom.

For each test, half of the questions were administered using a paper pencil test and half of the questions were administered using a computer based test. Each standard had an equal number of questions that were answered on paper and on the computer. Every question on the PPT had a DOK, depth of knowledge value, and there was a corresponding question on the CBT with an identical DOK. These questions were also identical in both construct and format. For example, if a standard required a student to solve a one-step equation, the student had a multiple-choice computation question on both the PPT and the CBT (Appendix A). If the question used a word problem format on the PPT, a corresponding word problem was found on the CBT. Both the PPT and the CBT

had flexibility in that students were permitted to skip test problems and return to them, as well as review their answers once they had completed the tests.

Student survey. All students were administered a 16-question survey, the 2017 Computer Access and Familiarity Survey from the National Assessment of Educational Progress, at the beginning of the study (Appendix B). This survey examined the access students have to various electronic devices, including Smart phones, tablets, laptops, and desktops, both at home and at school. It also explored the level of familiarity and comfort students had with technology and the variety of ways in which they interacted with technology for both personal and school use. The survey consisted of multiple choice, yes/no, and Likert scale items. Students were also asked to identify their gender on the survey. The survey was completed in approximately 25 minutes.

Procedures

The four classes that participated in the study were taught by a single 6th grade teacher, with the two co-taught classes having an additional teacher. All students participated in the instructional activities leading up the unit tests, including, but not limited to, direct instruction, individual practice, group practice, and formative assessment. Two of the classes, the co-taught classes, received instruction from the same two teachers. Each student completed the computer access and familiarity survey prior to taking the Unit 5 Test.

All students completed both unit tests. For each unit test, half of the questions were answered using a computer, and half of the questions were answered using a traditional paper-pencil test form. The Unit 5 test covers equations, and the Unit 6 test covers area and volume. The instruction for each of these units took approximately 4 weeks. All of the students received similar instruction and used similar instructional materials for both units. The test for each unit took place over the course of two days and signaled the completion of each 4-week unit. Unit 5 contains 3 standards; Unit 6 contains 3 standards. Each standard was assessed on both the paper-pencil portion and the computer portion using multiple choice, constructed response, and extended response items. This provided a comparison of test mode effect on different question constructs.

Students were placed in one of two conditions, either Condition 1: PPT first/CBT second or Condition 2: CBT first/PPT second. Each condition group contained forty students. In order to ensure equivalency between the two conditions, students were alternately placed in matched groups. These groups were determined based on the students' semester one course averages. The student with the highest course average from semester one was placed in Condition 1, and the student with the second highest course average was placed in Condition 2. This alternate placement of students into Condition 1 and Condition 2 continued until all students were placed.

Condition 1. The forty students in Condition 1 answered test questions from each standard on a paper pencil form and answered an equal number of questions from each standard on a computer based form. These questions included multiple choice, constructed response, and extended response questions. For each question on the PPT, there was a corresponding question on the CBT with a matching depth of knowledge, DOK, value. On the Unit 5 Test, Condition 1 students took the PPT first and the CBT second. On the Unit 6 Test, the order was reversed, so that Condition 1 students took the CBT first and the PPT second.

Condition 2. The forty students in Condition 2 answered test questions from each standard on a paper pencil form and answered an equal number of questions from each standard on a computer based form. These questions included multiple choice, constructed response, and extended response questions. For each question on the PPT, there was a corresponding question on the CBT with a matching depth of knowledge, DOK, value. On the Unit 5 Test, Condition 2 students took the CBT first and the PPT second. On the Unit 6 Test, the order was reversed, so that Condition 2 students took the PPT first and the CBT second.

Results

The purpose of this study was to determine if the testing format, PPT or CBT, had an effect on student performance on mathematical assessments. In order to test this, two unit tests were utilized, with students taking half of each test using PPT and the other half using CBT. The tests' mean scores were then compared.

Unit 5 Test

Prior to running analyses to test the main research questions, it was necessary to determine whether there was an effect from test order. To assess this, two independent samples t-tests were conducted. In the first, condition was the grouping variable and Unit 5 PPT was the dependent variable. Students in Condition 1 took this part first, and those in Condition 2 took it second. In the second independent samples t-test, condition was the grouping variable and Unit 5 CBT was the dependent variable. Students in Condition 2 took this part first, and those in Condition 1 took it second. While the order in which a student took a unit test, and its impact on student

achievement was not a specifically stated research goal, these tests were run to rule out the possibility that order influenced student achievement. There were no significant differences found between groups on either form, Unit 5 PPT p = .158 and Unit 5 CBT p = .374. This indicates that test order had no effect on student performance, as students performed similarly regardless of whether they took a certain form first or second.

The first research question sought to determine if there was a test mode effect on student test performance. To determine if there was a test mode effect, as evidenced by a difference in mean scores between the PPT and CBT, a paired sample t-test was used. Student scores on the Unit 5 PPT were higher, with a mean of 81.04, than on the Unit 5 CBT, with a mean score of 74.81. The mean score difference for the Unit 5 PPT and CBT was found to be statistically significant, $p < .001$. Descriptive and inferential statistic tables are shown below in Tables 1 and 2.

Table 1: Mean Score Difference Unit 5 PPT and CBT

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	PPT5	81.04	80	15.724	1.758
	CBT5	74.81	80	18.875	2.110

Table 2: Statistical Significance of Mean Score Difference Unit 5 PPT and CBT

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 PPT5 - CBT5	6.225	14.816	1.656	2.928	9.522	3.758	79	.000

The second research question was whether question type, multiple choice, constructed response, or extended response, had an effect on student performance on CBT compared to PPT. To determine this, additional paired sample t-tests were utilized. These tests examined the differences in mean scores for each of the three question constructs. The students had a mean score of 80.74 on the PPT multiple-choice questions in comparison to a mean score of 84.74 on the CBT multiple-choice questions. A statistically significant difference, p = .039, was found for the multiple choice mean score, in favor of the CBT. Descriptive and inferential statistic tables are shown below in Tables 3 and 4.

Table 3: Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MC5PPT	80.74	80	17.394	1.945
	MC5CBT	84.76	80	16.365	1.830

Table 4: Statistical Significance of Mean Score Difference Unit 5 PPT and CBT Multiple Choice

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 MC5PPT - MC5CBT	-4.025	17.190	1.922	-7.850	-.200	-2.094	79	.039

The difference in mean scores for the constructed response questions was also statistically significant, $p < .001$; however, in contrast to the multiple choice mean score difference, this was in favor of the PPT. Means and standard deviations are shown below in Tables 5 and 6.

Table 5: Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Constru5PPT	88.66	80	18.147	2.029
	Constru5CBT	65.00	80	29.926	3.346

Table 6: Statistical Significance of Mean Score Difference Unit 5 PPT and CBT Constructed Response

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1 Constru5PPT - Constru5CBT	23.662	29.747	3.326	17.043	30.282	7.115	79	.000

The third question construct, extended response, resulted in a mean score of 72.19 for PPT and 65.48 for CBT with a difference of 6.71 points in favor of the PPT, but this was not statistically significant $p = .099$.

To address the possibility of gender having an impact on the Unit 5 CBT scores, an independent samples t-test was run. The test resulted in a mean of 75.83 for females and 73.44 for males, with a mean difference of 2.38. This was not statistically significant, $p = .580$.

Finally, a Pearson Correlation was used to determine if a relationship existed between computer familiarity and CBT performance. Computer familiarity was determined by usage, or the amount of time each week a student spent using a computer for school work, including homework, and the frequency with which a student took tests on a computer. For CBT performance and computer usage, no statistically significant relationship was found, with $p = .690$. A correlational analysis of CBT performance and test frequency also resulted in no statistical significance, $p = .706$.

Unit 6 Test

Just as with the Unit 5 test, independent samples t-tests were run to determine whether test order, CBT or PPT first, had an effect on student achievement. Order was not found to produce a significant difference. For the PPT, the mean score difference of Condition 1 and Condition 2 was .25, $p = .941$. For the CBT, the mean score difference of Condition 1 and Condition 2 was 3.5, $p = .437$. This again demonstrated that the order in which a student took the test, PPT first followed by CBT or CBT first followed by PPT, did not have an effect on performance.

To answer the first question of whether a test mode effect existed, as evidenced by a difference in mean scores between the PPT and CBT, a paired samples t-test was used. The mean scores for the Unit 6 PPT and CBT were 80.13 and 77.63, respectively, with a mean difference of 2.5, $p = .224$, indicating there was not a significant difference between the means for the PPT and the CBT.

To address the question of whether question construct had an effect on student performance, additional paired samples t-tests were implemented to determine the mean score differences for each of the three question constructs. In contrast to the Unit 5 multiple-choice mean scores, the mean score of the Unit 6 PPT multiple choice mean was higher, 88.75, than that of the Unit 6 CBT multiple choice, 82.49. A significant difference, $p = .002$, was found for the multiple choice mean score differences, indicating that student performance was greater on PPT constructed response. Tables 7 and 8 show the descriptive and inferential statistics.

Table 7: Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	MC6PPT	88.75	80	13.069	1.461
	MC6CBT	82.49	80	19.596	2.191

Table 8: Statistical Significance of Mean Score Difference Unit 6 PPT and CBT Multiple Choice

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1 MC6PPT - MC6CBT	6.263	17.056	1.907	2.467	10.058	3.284	79	.002

The mean score for the constructed response questions on the Unit 6 PPT was 80.94, and the mean score for the constructed response questions on the Unit 6 CBT was 80.00, resulting in no significant difference, $p = .831$. Though the mean scores for the extended response had greater variation, 54.56 for the PPT and 62.50 for the CBT, with a mean difference of 7.93, this was not statistically significant, $p = .070$.

To address the research question of gender having an impact on the Unit 6 CBT scores, an independent samples t-test was run. The test resulted in a mean of 77.83 for females and 77.35 for males, with a mean difference of .47. This was not statistically significant, $p = .917$.

The final research question was whether a relationship existed between computer familiarity and CBT performance. Computer familiarity was determined by usage, or the amount of time each week a student spent using a computer for school work, including homework, and the frequency with which a student took tests on a computer. Pearson Correlations were used to test for a relationship between computer familiarity and CBT performance. For CBT performance and computer usage, no significant relationship emerged, with $p = .446$. Similarly, for CBT performance and test frequency, the correlational analysis also revealed no significant relationship, $p = .096$. These findings suggest that computer familiarity, as measured by computer usage and test frequency, did not have an impact on student test performance.

Discussion

This study aimed to determine if a test mode effect existed for student performance on paper pencil tests and computer based tests. The study also sought to determine if the question construct had an effect on student performance on a CBT versus a PPT. Finally, the impact of either gender or computer familiarity in relation to CBT performance was considered.

In regard to the central question of test mode effect, results for overall test mode varied from the Unit 5 test to the Unit 6 test. The difference in mean scores was only found to be statistically significant for the Unit 5 test, in favor of PPT. While this indicates a possible benefit for students taking a traditional paper-pencil mathematics test, the mean score difference for the Unit 6 test did not corroborate this finding. The Unit 5 test, which resulted in a statistically significant mean score difference favoring PPT, tested students on algebraic expressions and equations. The Unit 6 test assessed students on area and volume, a part of the geometry domain. These results are consistent with Logan's (2014) results in which data from sixth grade students' math scores were found to be statistically significant in favor of PPT in the algebra domain, but not the geometry domain. While Keng (2008) posited that higher student achievement on PPT may be due to students' ability to draw on the item or geometric figure, the results of this study were found to be in contrast, as there was no significant difference in PPT and CBT scores on the Unit 6 test that assessed the geometry domain.

The type of question, multiple choice, constructed response, or extended response, was also evaluated for a possible effect on student performance. The mean score differences on the multiple-choice items were statistically significant for both unit tests, with the Unit 5 multiple choice mean favoring the CBT and the Unit 6 multiple choice mean favoring the PPT. The Unit 6 test assessed students on problems from the geometry domain. Keng (2008) posited that the ability to correctly solve geometry problems is made easier with the ability to draw on the shapes on the paper version, as opposed to having to correctly transfer the drawing from a computer based test to paper. The higher score on the Unit 6 PPT multiple choice could be attributed to the students' ability to interact with the geometric shapes on paper. Looking at both tests together, though, the fact that one particular test mode did not consistently yield higher scores for multiple choice questions suggests that multiple choice items are not subject to mode effect. Testing mode, either PPT or CBT, did not consistently give students an advantage on multiple choice items.

Constructed response achievement favored the PPT on the Unit 5 test, but not the Unit 6 test. It has been suggested that due to the limitations imposed by CBTs, students experience increased success on constructed and extended response on PPTs, as PPTs more readily allow students to include pictures or diagrams to support their answers (Bennett et al., 2008). For the Unit 5 Test on equations, students in the study were able to employ a problem-solving strategy for one step equations on the PPT that may have led to increased success as compared to the CBT. While students taking the CBT did have access to scratch paper, it is possible they did not employ the same problem-solving strategy for one step equations because this involved an extra step of transferring from the scratch paper to the computer. Students taking a CBT may consciously make the choice of not going to the effort of using additional material, such as scratch paper, beyond the CBT, or it is possible that they may become so engrossed with the computer itself, that they do not remember to use the scratch paper.

For both Unit 5 and Unit 6, student performance on the CBT and PPT extended response was comparable. Charman's (2014) results indicate that when students are answering extended result essay questions they create longer, more detailed answers on a computer than on paper. This would seem to suggest that scores would have been higher on the CBT, but this was not the case for this study. Though differences did exist in the mean scores for both unit tests, they were not statistically significant. This could be due in part to the large standard deviations or to the limited sample size of only 80 students. Scores may also have been comparable for the PPT and CBT extended response because the very nature of an extended response question allows a student to express their own thought process and support it with details, as opposed to a multiple-choice question which has only one specific answer.

When the impact of gender on CBT scores was considered in this study, no relationship was found to exist. Though this is not consistent with Jeong's (2012) study which resulted in higher test scores for males than females on CBTs, this is consistent with several other studies that have not found an association between gender and CBT performance (Bennett et al., 2008; Lee et al., 2010; Poggio et al., 2005). With gender not having an effect on student CBT performance in this study, and with the fact that this was consistent with previous studies, it is possible that a gender gap no longer exists for computer competency. This may be due to several factors, including the ubiquity of technology in the homes and of computer literacy courses in school.

Results also demonstrated that there was also not a relationship between computer familiarity and CBT scores. Familiarity was determined by student responses on the 2017 National Assessment of Educational Progress Student Survey Questionnaires: Computer Access and Familiarity Study Grade 4. Specifically, familiarity was determined by student computer usage for classwork, including homework, and by the frequency which students had taken computer tests during the current school year. This suggests that while the amount of time a student spends using a computer may vary, it does not have an impact on test performance, as all students had a base level of computer familiarity.

Limitations

Several points can be noted and can be considered limitations to the study. First, the participants in this study came only from one teacher's classroom. Before generalizations can be made, it would be necessary to widen the scope of the classes tested. Other teachers may use computers more or less frequently in their classrooms for daily assignments or for tests, resulting in students having a greater or lesser level of comfort with CBTs. Also, the students in this study are in a school district that has a long history of BYOT, bring your own technology, so the participants had a high level of computer familiarity in an academic setting. This familiarity could vary in other school districts.

Another consideration of the study is the number of each type of question on the unit tests. While both unit tests had six multiple choice questions on both the PPT and CBT, there was only one to two constructed and extended response questions on each test mode. The length of time required for students to fully answer constructed response and extended response questions was the reason for the abbreviated number of these types of question; however, with so few questions, a limited picture of student performance on these question constructs exists.

Future Research/Implications

Despite the limitations, the study resulted in varied student achievement for students on PPT as compared to CBT in two different mathematical domains and question constructs. Although the reasons for this are unclear, these results suggest that further inquiry into test mode effect and question construct effect is warranted. Additional research could also examine the varying requirements for solving problems in the different mathematical domains, such as algebra, geometry, and statistics.

The study's results indicate that gender did not have an impact on student CBT performance, nor did computer familiarity. Future research could examine other possible factors, such as overall student academic achievement, student preference, and student knowledge of and ability to use computer tools, and their possible impact on CBT performance.

Conclusions

Given the integral role of technology in society today, computers as instructional and assessment tools in schools currently are both widely accepted and practical. The benefits of using computers for testing are numerous. Therefore, efforts must be made to ensure student performance on CBTs is an accurate indicator of content competency. The consistent utilization of computers for student classwork and homework will allow the student to be as comfortable with a CBT as with a PPT. Additionally, technological guidance must be provided to

students on the range of tools, such as equation editor, annotation tools, and geometry manipulation, to support student CBT achievement.

References

Bayazit, A., & Askar, P. (2012). Performance and duration differences between online and paper-pencil tests. *Asia Pacific Education Review, 13*(2), 219-226. doi:10.1007/s12564-011-9190-9

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, And Assessment*, *6*(9), 1-38.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51-60.

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.

Charman, M. (2014). Linguistic analysis of extended answers: Differences between on-screen and paper-based, high- and low-scoring answers. *British Journal of Educational Technology, 45*(3), 834-843. doi: 10.1111/bjet.12100

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Testing, 33*(5), 595-604. doi:10.111/1467-8535.00294

Dimcock, P.H. (1991). The effects of format differences and computer experience on performance and anxiety on a computer administered test. *Measurement & Evaluation in Counseling & Development, 24*(3), 1-8.

Eid, G. K. (2005). An investigation into the effects and factors influencing computer-based online math problem-solving in primary schools. *Journal of Educational Technology Systems*, *33*(3), 223-240.

Flowers, C., Kim, D., Lewis, P., & Davis, V. C. (2011). A comparison of computer-based testing and pencil-and-paper testing for students with a read-aloud accommodation. *Journal of Special Education Technology*, *26*(1), 1-12.

Jeong, H. (2012). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. doi:10.1080/0144929x.2012.710647

Johnson, M., & Green, S. (2006). On-line mathematics assessment: The impact of mode on performance and question answering strategies. *Journal of Technology, Learning, And Assessment*, *4*(5), 1-33.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education*, *21*(3), 207-226. doi:10.1080/08957340802161774

Kim, D., & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *Journal of Technology, Learning, and Assessment, 6*(4), 1-30.

Lee, K. S., Osborne, R. E., & Carpenter, D. N. (2010). Testing accommodations for university students with AD/HD: Computerized vs. paper-pencil/regular vs. extended time. *Journal of Educational Computing Research*, *42*(4), 443-458. doi:10.2190/EC.42.4.e

Logan, T. (2015). The influence of test mode and visuospatial ability on mathematics assessment performance. *Mathematics Education Research Journal, 27*(4), 423-441. doi:10.1007/s13394-015-0143-1

Minnesota Department of Education. (2012). *Mathematics Minnesota comprehensive assessment-series III: Mode comparability study report.* Retrieved from http://www.education.mn.gov/MDE/dse/test/mn/Tech/index.html

National Assessment of Educational Progress. (2017) Student Survey Questionnaires: Computer Access and Familiarity Study Grades 4 & 8. Retrieved from https://nces.ed.gov/nationsreportcard/subject/field_pubs/sqb/pdf/2017_sq_computer_access_familiarity.pdf

Piatt, C., Coret, M., Choi, M., Volden, J., & Bisanz, J. (2016). Comparing children's performance on and preference for a number-line estimation task: Tablet versus paper and pencil. *Journal of Psychoeducational Assessment, 34*(3), 244-255.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, And Assessment*, *3*(6), 1-30.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining differences in examinee performance in paper and pencil and computerized testing. *Journal of Technology, Learning, And Assessment*, *6*(3), 1-20.

Taherbhai, H., Seo, D., & Bowman, T., (2012). Comparison of paper-pencil and online performances of students with learning disabilities. *British Educational Research Journal, 38*(1), 61-74. doi: 10.1080/01411926.2010.526193

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, *66*(3), 335-348. doi:10.1007/s10649-006-9078-5

United States Census Bureau. (2016). *Quick facts Cumming City, GA* [Data file]. Retrieved from https://www.census.gov/quickfacts/table/PST045215/1320932,13117,00

