

---

Theses and Dissertations

---

Spring 2015

# Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement

Kiersten Kenning Hensley  
*University of Iowa*

Copyright 2015 Kiersten Kenning Hensley

This dissertation is available at Iowa Research Online: <http://ir.uiowa.edu/etd/1627>

---

## Recommended Citation

Hensley, Kiersten Kenning. "Examining the effects of paper-based and computer-based modes of assessment on mathematics curriculum-based measurement." PhD (Doctor of Philosophy) thesis, University of Iowa, 2015.  
<http://ir.uiowa.edu/etd/1627>.

---

Follow this and additional works at: <http://ir.uiowa.edu/etd>



Part of the [Teacher Education and Professional Development Commons](#)

EXAMINING THE EFFECTS OF PAPER-BASED AND COMPUTER-BASED  
MODES OF ASSESSMENT ON MATHEMATICS CURRICULUM-BASED  
MEASUREMENT

by

Kiersten Kenning Hensley

A thesis submitted in partial fulfillment  
of the requirements for the Doctor of Philosophy  
degree in Teaching and Learning (Special Education)  
in the Graduate College of  
The University of Iowa

May 2015

Thesis Supervisor: Professor John L. Hosp

Copyright by

**KIERSTEN KENNING HENSLEY**

2015

All Rights Reserved

Graduate College  
The University of Iowa  
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

PH.D. THESIS

---

This is to certify that the Ph.D. thesis of

Kiersten Kenning Hensley

has been approved by the Examining Committee for  
the thesis requirement for the Doctor of Philosophy degree  
in Teaching and Learning (Special Education) at the May 2015 graduation.

Thesis Committee:

\_\_\_\_\_  
John L. Hosp, Thesis Supervisor

\_\_\_\_\_  
Allison Bruhn

\_\_\_\_\_  
Anne Foegen

\_\_\_\_\_  
Kristen Missall

\_\_\_\_\_  
Suzanne Woods-Groves

To my husband and best friend, Shane, for supporting me in this crazy life

## **ACKNOWLEDGEMENTS**

In a meeting eight years ago, a colleague told me I should go get my Ph.D. I laughed, feeling like it was impossible at the time. Thanks to my colleagues at the Iowa Department of Education for their encouragement and support. In particular, I would like to thank Dr. Michelle Hosp and Dr. Barbara Guy for giving me the push I needed to commit to the process.

To those who served on my committee, thank you for your time and assistance with my research project. Dr. John Hosp, my advisor and dissertation chair, thank you for keeping things humorous as you provided mentorship, guidance, and feedback. Thank you to Dr. Anne Foegen for providing me with many opportunities. To Dr. Allison Bruhn, Dr. Kristen Missall, and Dr. Suzanne Woods-Groves, thank you for your time and feedback, it is truly appreciated.

To my family, thank you for your love and support. Special thanks to my mother, Dr. Phyllis Anderson, for putting up with me when I made a mess of your house and fixing eggs for me every morning, and for being a role model for how to stay connected to the real needs of classroom teachers and students. To my husband, Shane, I can never thank you enough for putting up with me, never complaining when I spent half of the last four years away from home, and for helping me with all my tech needs. And to Ruby, the best dog ever, thank you for always sitting with me as I read and write.

To my cohort members, classmates, and friends made through this process, this wouldn't have been near as much fun without you! Special thanks to Kari, Sally, Jeremy, and Kris for your support.

## ABSTRACT

The computer to pupil ratio has changed drastically in the past decades, from 125:1 in 1983 to less than 2:1 in 2009 (Gray, Thomas, & Lewis, 2010), allowing for teachers and students to integrate technology throughout the educational experience. The area of educational assessment has adapted to the increased use of technology. Trends in assessment and technology include a movement from paper-based to computer-based testing for all types of assessments, from large-scale assessments to teacher-created classroom tests. Computer-based testing comes with many benefits when compared to paper-based testing, but it is necessary to determine if results are comparable, especially in situations where computer-based and paper-based tests can be used interchangeably.

The main purpose of this study was to expand upon the base of research comparing paper-based and computer-based testing, specifically with elementary students and mathematical fluency. The study was designed to answer the following research questions: (1) Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer? (2) Are there differential mode effects on computer-based tests based on sex, grade level, or ability level?

A mixed-factorial design with both within- and between-subject variables was used to investigate the differences between performance on paper-based and computer-based tests of mathematical fluency. Participants completed both paper- and computer-based tests, as well as the Group Math Assessment and Diagnostic Evaluation as a measure of general math ability. Overall findings indicate that performance on paper- and computer-based tests of mathematical fluency are not comparable and student grade-level may be a contributing factor in that difference.

## **PUBLIC ABSTRACT**

With the increase in available classroom technology, it is necessary to examine how assessments administered through the use of technology compare to traditional, non-technology based methods. Paper-based and computer-based tests are often used interchangeably, depending on teacher and student comfort level with technology, therefore it is necessary to examine how paper- and computer-based tests compare.

This study examined the comparability of paper-based and computer-based tests designed to measure mathematical fluency, as well as how the factors of age, sex, and general ability in mathematics contributed to the comparability of paper-based and computer-based tests. Results showed that scores obtained through the administration of identical paper-based and computer-based tests of mathematical fluency are not comparable. There was no difference in how males and females performed, but the computer-based tests had a greater effect on fifth grade students than fourth grade students. There was also not a consistent relationship between the level of overall mathematical ability and performance on computer-based tests.



## TABLE OF CONTENTS

LIST OF TABLES .....	viii
LIST OF FIGURES .....	ix
CHAPTER ONE. INTRODUCTION.....	1
Curriculum-Based Measurement.....	2
Mode Effects of Assessment .....	4
Purpose of the Study .....	6
Importance of the Study .....	6
CHAPTER TWO. LITERATURE REVIEW .....	7
Technology in Schools .....	7
Technology for Teaching and Learning .....	8
How Technology has Changed Assessment .....	9
Technology and Large Scale Assessments .....	10
Technology and Classroom Assessments .....	14
Curriculum-Based Measurement.....	15
Technology and CBM .....	19
Mode Effects and Assessment.....	21
Technology Issues .....	22
Participant Factors .....	25
Issues in Mode Assessment Research .....	33
Overview and Research Questions.....	34
CHAPTER THREE. METHOD .....	36
Overview.....	36
Participants and Settings.....	36
Study Design.....	37
Instruments .....	38
Procedures .....	40
Analysis of Results .....	43
CHAPTER FOUR. RESULTS .....	45
Findings for Research Question 1 .....	45
Comparability Results .....	45
Comparable Reliabilities .....	46
Correlate Between Test Modes.....	48
Correlate with Criterion Measure .....	53
Comparable Means and Standard Deviations.....	53
Findings for Research Question 2 .....	54

Sex .....	55
Grade .....	55
Overall Mathematics.....	56
Summary of Results .....	59
CHAPTER FIVE. DISCUSSION.....	60
Research Question 1 .....	60
Research Question 2.....	64
Limitations .....	66
Implications for Practice .....	67
Future Research.....	68
Conclusion.....	69
APPENDIX A. DIRECTIONS FOR PAPER-BASED PROBES .....	71
APPENDIX B. PAPER-BASED PROBE 1 (PAGE 1 OF 3).....	72
APPENDIX C. DIRECTIONS FOR COMPUTER-BASED PROBES .....	73
APPENDIX D. COMPUTER-BASED PROBE 1 (PAGE 1 OF 3).....	74
APPENDIX E. FOURTH GRADE PARTICIPANT RANKING .....	75
APPENDIX F. FIFTH GRADE PARTICIPANT RANKING .....	78
REFERENCES .....	81

## LIST OF TABLES

Table 1 Summary of Studies on Mode Effects of Assessment.....	31
Table 2 Participant Demographics.....	37
Table 3 Reliability of Pilot Study Measures.....	40
Table 4 The Distribution of Problems Correct on PBT and CBT M-CBM Probes.....	54
Table 5 Descriptive Statistics for Performance of Male and Female Participants Across PBT and CBT Modes of Assessment .....	55

## LIST OF FIGURES

Figure 1 Welcome Screen for CBT Probes.....	41
Figure 2 CBT Sample Test .....	42
Figure 3 Parallel-Forms Reliability PBT Forms.....	47
Figure 4 Parallel-Forms Reliability CBT Forms .....	47
Figure 5 Scatterplot of PBT and CBT Forms of M-CBM Probes- All Participants .....	48
Figure 6 Scatterplot of PBT and CBT Forms of M-CBM Probes- Fourth Grade.....	49
Figure 7 Scatterplot of PBT and CBT Forms of M-CBM Probes- Fifth Grade .....	49
Figure 8 Ranking of Individuals on PBT and CBT Modes- Fourth Grade.....	51
Figure 9 Ranking of Individuals on PBT and CBT Modes- Fifth Grade.....	52
Figure 10 Quantile Plots .....	58

## **CHAPTER ONE**

### **INTRODUCTION**

In 1983, the ratio of computers to students in K-12 classrooms was 125:1. In the last 30 years, the ratio has dropped drastically. In 1995, the ratio of computers to students was 9:1, 6:1 in 1998, 4:1 in 2002, and 1.6:1 in 2009 (Project Tomorrow, 2014). In instructional settings, the availability of technology allows for greater mobility for students, teachers, and parents through collaboration and communication that goes beyond the school walls (Holland & Holland, 2014). Lessons can be built on vivid multi-media experiences (Reiser & Dempsey, 2007) and increased ability to provide varied instruction and access to differentiated curricular materials to meet the needs of diverse learners in the classroom (Holland & Holland, 2014; Nash, 2009). One-to-one initiatives have put laptops and tablets in the hands of students in over half of the districts in Iowa alone (“1:1 Schools in Iowa,” 2012). The availability of technology in schools is supported by both private and public funding. In 2012, the Bill and Melinda Gates Foundation provided more than 20 million dollars in grants to schools for innovative practices in using technology to improve educational outcomes (Gates Foundation, 2014). Digital Promise is a government and private collaboration, designed to encourage technological innovation in education and close the digital learning gap due to a lack of equity in technological resources (Digital Promise, 2014)

With more teachers using technology in their classrooms to support instruction, it is expected that the shift to technology use would also happen with assessment. Computer-based tests (CBT) include assessments that are completed using the computer, either through a computer program or through a web-based system. Many assessments

have already been converted into a computerized format, including large scale, summative, and daily formative assessments (Pellegrino & Quellmalz, 2010). There are some obvious advantages to CBTs, including increased student motivation, improved accuracy in data collection, improved match for special populations, and fast reporting of results (Kapoor & Welch, 2011; Poggio & McJunkin, 2012). Studies have also found that students enjoy taking CBTs and are motivated by the use of technology (Bodmann & Robinson, 2004; Ripley, 2009). Accessibility features can be conveniently embedded into CBTs, allowing for supports and accommodations such as glossaries, color contrast, text-to-speech, spell check, highlighting, and closed-captioning (Bennett, 2015). The quick reporting of results is particularly useful to educators as they are able to access and use these results to make timely changes in instruction if necessary (Bennett, 2003; Dean & Martineau, 2012; Peak, 2005; Poggio & McJunkin, 2012).

### **Curriculum-Based Measurement**

One quick, valid, and reliable way to collect information regarding student progress in the curriculum is through the use of curriculum-based measurement (CBM). CBM uses concise and uncomplicated measures to assess skill and progress in the basic academic areas of reading, writing, and mathematics, supporting teachers in making timely instructional decisions (Stecker, Lembke, & Foegen, 2008) for students from early childhood, elementary, and secondary levels (Deno, 2003).

Traditionally, CBM probes have been administered using paper-based tests (PBT), meaning that pages of problems were given to the student for him or her to complete using a writing utensil. Numerous studies have found high levels of reliability and validity using paper and pencil CBM (Fuchs, Fuchs, Hamlett, & Stecker, 1991;

Lembke, Hampton, & Beyers, 2012). These measures have also been found to be repeatable, sensitive to student growth over time, and helpful to educators in making decisions about what to teach (Shapiro, 2004).

With all of the benefits of CBM, there are some issues that arise in the use of paper and pencil CBM, including printing costs, administration time, and consistent scoring. The costs that go with printing weekly CBM measures can add up (e.g., math CBM measures are five to seven pages per student per week). While CBM is meant to be a quick and concise measure, it still requires a time commitment in administration, scoring, and data entry. Any individual CBM measure takes between one and eight minutes to administer, not all are administered by group, and this does not include scoring each individual student (Deno, 2003). Another area of concern is consistent scoring. Measures scored by teachers or paraprofessionals can lead to inconsistent or incorrect scoring practices (Fuchs, Fuchs, & Hamlett, 1994).

In an attempt to alleviate these issues, some CBM measures have been combined with technology. Numerous assessment suites have created computer-based assessments (AIMSweb, Data Director, easyCBM, mCLASS, FastBridge, and Yearly Progress Pro). There are many advantages for educators using these computer-based CBM, including automatic scoring, options of displays, and immediate logging of scores (Bridgeman, 2009; Redecker & Johannessen, 2013), all of which are not options for paper-based CBM.

Some concerns arise when considering the use of CBTs for CBM. It is important to determine if PBT and CBT formats of CBM are equivalent. In order to be considered equivalent, scores and distributions need to be similar across both formats (Pomplun,

Frey, & Becker, 2002). The research on CBM and equivalency across PBT and CBT formats is lacking. Researchers have been comparing PBT and CBT on other types of assessments for decades and trying to understand why there may be a difference in performance across tools, even when the PBT and CBT are identical, but this research has not happened with CBM.

### **Mode Effects of Assessment**

When PBT and CBT formats of the same assessment produce different results, this performance difference is called a mode effect (Pomerich, 2004). Standard 4.10 of the *Standards for Educational and Psychological Testing* specifies that before a new mode of assessment is used, empirical evidence must show that the change in mode does not put any students at a disadvantage (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In order to test for these differences, studies must be conducted to compare the modes of assessment. The International Testing Commission (2006) provides a framework for examining the comparability of assessments administered through PBT and CBT modes. The guidelines include a wide variety of factors to consider for test developers, test publishers, and test users, including four statistical characteristics that help ensure equivalence between formats:

- Comparable reliabilities
- Adequate correlations between formats
- Comparable correlation with external criterion measure
- Comparable means and standard deviations



If PBT and CBT formats show comparability as determined by the ITC guidelines, then it can be assumed that the formats are comparable. However, if it has been determined that the scores on two identical PBT and CBT assessments are not comparable, then it is necessary to examine why those mode effects exist. There are two main categories in the research on examining mode effects: technology issues and participant factors. Technology issues could impact performance through issues of screen resolution, font size, lack of reliable tools, or even the ability to review and change answers (Gould, Alfaro, Finn, Haupt, & Minuto, 1987; Vispoel, 2000). Characteristics of the test taker could also impact the comparability of PBT and CBT scores. Characteristics that have been studied include sex, ethnicity, familiarity with technology, socio-economic status, and ability in the content area (Kolen, 1999; Pomerich, 2004; Wheadon, 2011). Studies have also examined how characteristics of the test format could sway score comparability, including whether or not answers can be changed, size and color of font, and whether or not the test is timed (Wheadon, 2011).

Results from research in this area is mixed. Some studies show that administration mode has little or no effect on student performance (Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004; Peak, 2005; Wang, Jiao, Young, Brooks, & Olson, 2007), while others have found CBT items to be more difficult for students than PBT items (Bennett, 2003; Choi & Tinkler, 2002). Other researchers found CBT performance to be higher than PBT performance (Bugbee & Bernt, 1990; Clariana & Wallace, 2002; Pomplun, Frey, & Becker, 2002). Many of these studies will be discussed in detail in Chapter 2.

### **Purpose of the Study**

This study was designed to determine if there is a mode effect between paper- and computer-based CBM measures of math computation for fourth and fifth grade students. Participants completed identical paper-based and computer-based CBM math fact probes. This study also examined possible factors that can lead to mode effects, including sex, grade, and ability in the content area. The following research questions were addressed:

- 1) Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer?
- 2) Are there differential mode effects based on sex, grade level, or ability level?

### **Importance of the Study**

Even though mode effects of different formats of assessments have been studied for decades, it is important to continue the line of research. As the availability of technology in classrooms increases, and as teachers become more technologically savvy, it is essential they understand how switching from paper-based to computer-based tests may have an impact on student performance. This is of particular importance when teachers have the ability to create their own computer-based versions of assessments that are used between students and classrooms and are used to make instructional decisions.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

It is impossible to predict the future, but if recent movements in technology in education persist, technology will continue to impact education in powerful ways, including how students are assessed. This chapter includes an overview of research on technology and assessment, particularly how technology has changed assessment practices in schools. Beginning with research on technology and large-scale assessments, this leads into how that information has translated into examinations on the use of technology with classroom-based assessments. This leads to background and studies on curriculum-based measurement, which is one type of classroom based assessment. The next section of the literature review includes an evaluation of studies on mode effects related to PBT and CBT formats, organized around issues in technology and participant factors. The last section details some of the issues prevalent in mode assessment research.

#### **Technology in Schools**

Technology use in schools has been steadily on the rise. One-third of K-12 students in the U. S. have access to school-issued mobile devices. Eighty-nine percent of high school students and 73 percent of middle schools students use smart phones, with another 66 percent of middle and high schools students reporting access to laptops (Project Tomorrow, 2014). In a survey of educational technology in U.S. public schools, Gray, Thomas, and Lewis (2010) found the ratio of students in the classroom to computers (in or can be brought into the classroom) is 1.6 to 1 at the elementary and 1.7 to 1 at the secondary level. Low ratios of computers to students can go a long way to increase access to technology, but does it change how and how often technology is used?

## **Technology for Teaching and Learning**

Technology can take on numerous functions in the classroom, such as serving as a resource for more information, a conduit for the delivery of instruction, a tool for motivation, a means for communication, and an assessment and data collection instrument (Lee, Waxman, Wu, Michko, & Lin, 2013).

One-to-one initiatives are increasing in popularity in K-12 classrooms. Four percent of U. S. school districts implemented 1:1 programs in 2003. In 2006, the estimate of 1:1 programs was closer to 25% (Bebell & Kay, 2010). One-to-one programs ensure that each student has access to his/her own device and does not have to share. Research on 1:1 initiatives has shown they are effective at increasing student engagement (Penuel, 2006), decreasing discipline issues (Baldwin, 1999), and increasing time engaged in schoolwork outside of school (Russell, Bebell, & Higgins, 2004).

Not all schools are able to provide a laptop or tablet to each student, but the increased use of technology in classrooms in general has shown positive effects. Harrison, Lunzer, Tymms, Taylor, and Restorick (2004) studied how low, mid, and high levels of classroom technology correlated with England's national assessments. At all grade levels, they found a significant advantage associated with higher levels of classroom technology and English, mathematics, science, and foreign languages. In a meta-analysis covering 40 years of studies on classroom technology, Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011) found the research to support positive outcomes from the use of classroom technology.

In many ways, technology has changed the traditional learning environment. Different types of technology, from video games to social networks to hand held tablets

and smart phones have changed the way we teach and learn (Halvorson & Shapiro, 2012) This change in the way students are taught should also align with how students are assessed.

### **How Technology has Changed Assessment**

Reported benefits of CBT include easier administration, lower printing costs and paperless distribution, increased test security, rapid feedback, and standardized support tools (Bridgeman, 2009). CBTs can easily manage test timing, ensuring that all students are given the same amount of time when completing a test. When the test is in the format of an electronic file, it is no longer necessary to print, keep track of, and mail. When test security is an issue, CBT files can be sent just before the testing window, decreasing the possibility of test questions being uncovered before the test. CBTs can be instantly scored, providing usable feedback to students and teachers. CBTs can also allow for students to have access to the same support tools, including dictionaries, calculators, text-to-speech, and also only allow the use of those tools when it is appropriate (Bridgeman, 2009).

With all of these benefits, it is easy to see why many schools are employing the use of technology for assessments. Forty-four percent of students in upper elementary school are taking tests online. In 2013, over 50 percent of middle and high school students are taking tests online. This is an increase from 2009, when it was reported that 39 percent of secondary students were taking tests online (Project Tomorrow, 2014). This level of use falls in line with estimates found by Grey, Thomas, and Lewis (2010) with 84% of U.S. K-12 teachers reporting the availability of technology to administer assessments, with 61% of those teachers reporting that they use technology to administer

assessments often. This estimate includes more technology-based options than just online assessments, such as classroom response systems and computer-based assessments. Sixty-four percent of elementary teachers reported using technology-based assessments often, as opposed to only 55% of secondary teachers (Grey, Thomas, & Lewis, 2010). With such large numbers of schools using technology-based assessments, it is more than likely some of them are large-scale assessments.

### **Technology and Large Scale Assessments**

Large-scale assessments test large numbers of students, and include national and international assessments, statewide testing programs, or assessments systems used by districts. Historically, the purpose of large-scale assessments has been to provide schools, districts, and states information on accountability related to learning standards and reform efforts (Taylor, 1994). Large-scale assessments are often administered at the command of users outside the classroom, such as policy makers at the district, state, and national level (Chudowsky & Pellegrino, 2003). Therefore, the intent of large-scale assessments is to generalize across a wide variety of contexts, leading to the ability to make comparisons across schools, districts, states, and even countries, regardless of differences in classroom and educational experiences (Brookhart, 2003).

Technology in assessment has been steadily implemented in the area of large-scale assessments. In the initial phases of CBTs for large-scale assessments, many were exact replications of the PBT. Early use of technology with large-scale assessments focused on efficiencies, both logistical and financial (Pellegrino & Quellmalz, 2010). Paperless online tests eliminate the need for printing, shipping, and keeping track of paper testing materials. Data collection and scoring requires little work or wait on the part

of teachers, as computer scoring provides almost instantaneous results. Computer scoring also leads to quick feedback that can be customized to a variety of audiences (Quellmalz & Pellegrino, 2009). All of these options make CBT an attractive choice, even when the CBT assessment is simply a direct translation of the PBT (Becker, 2006; Redecker & Johannessen, 2013), but recent advances in technology have built upon that initial focus on efficacy and have expanded the possibilities for CBT and large-scale assessments.

Developments in technology now allow for the capture of detail that was not possible with paper-based assessments. Problem-solving sequences can be monitored, allowing for the examination of how a student arrives at his or her answer. This can provide information to educators on where mistakes or misunderstandings occur, leading to the ability to make stronger instructional decisions, and also allow for the extraction of patterns that correlate with levels of achievement (Quellmalz & Pellegrino, 2009; Pellegrino & Quellmalz, 2010). Monitoring the use of time is also possible with some large-scale CBTs, which can give educators more insight to the student and his/her instructional needs by analyzing sections where students have to spend more time and effort (Pellegrino & Quellmalz, 2010). Advancements in CBT scoring now allow for the computerized scoring of essays and constructed response items through techniques such as latent semantic analysis (Quellmalz & Pellegrino, 2009). Such techniques have been found to produce comparable human and machine-graded scores (Klein, 2008).

Advances in assessment technology do not just benefit administration and scoring, but also allow for a different experience for the test-takers. Instead of multiple-choice questions, CBTs can more easily include assessments with embedded multimedia, leading to simulated experiences that test-takers can interact with and respond to

(Quellmalz & Pellegrino, 2009). CBTs also allow for built-in options for universal design and assistive technology principles. Universal design includes options for presentation and representation of questions and concepts, allowing for more test-takers to be assessed fairly (Beller, 2013). Embedded assistive technology options include accommodations for test-takers with special needs, such as American Sign Language, Braille, closed captioning, and text-to-speech, allowing persons with disabilities to demonstrate understanding independently (Beller, 2013). In a PBT, students with disabilities would have to depend on someone to provide those accommodations, but when the options are embedded into a CBT, they can be independent.

Examples of large scale CBTs employing the use of advanced technology include the Programme for International Student Assessment (PISA; OECD, 2007), the National Assessment of Educational Progress (NAEP; NCES, 2011), Smarter Balanced assessments (SBAC, 2014), and Partnership for Assessment of Readiness for College and Careers (PARCC, 2013) assessments. The PISA began with science CBTs in 2006, with the purpose of not just testing knowledge, but also testing the process of inquiry, which had not been captured in previous PBT versions (Quellmalz & Pellegrino, 2009). Computer-based PISA assessments are now in use in the areas of reading, mathematics, and problem solving, and involve the use of interactive multimedia materials (OECD, 2007). The NAEP assessment began using interactive computer tasks in the form of simulated experiments and investigations with science assessments in 2009, and the NAEP writing assessment began requiring word processing and editing tools in 2011 (NCES, 2011). Both SBAC and PARCC have implemented principles of universal design and assistive technology to make CBTs more accessible for all students by building in a



long list of accessibility options including multiple languages, text-to-speech, increased font sizes, dictionaries, highlighters, American Sign Language, and Braille (PARCC, 2013; SBAC, 2014).

In the early 2000s, CBT for large-scale testing in individual states was just in the development phases, but quickly became an established expectation (Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Thurlow, Lazarus, Albus, & Hodgson, 2010). The switch from PBT to CBT was also encouraged by the U. S. Department of Education's Race to the Top initiative, with an obvious push to use technology for assessment, including millions of dollars allocated to PARCC and SBAC to create computer-based assessments (Fletcher, 2010). Given their use and encouragement to switch from PBT to CBT it is necessary to consider concerns there might be when using CBT for large-scale assessments. These concerns include disadvantages for some students due to issues with access, unequal availability of technology, and lack of infrastructure to support online assessments (Thurlow et al., 2010). Many schools do not have enough computers to allow for entire classes to test at the same time (Thurlow et al., 2010). Even if they do have enough computers, many schools do not have the level of bandwidth necessary for large numbers of children to take assessments online at the same time (Beller, 2013).

Even with these issues with computer-based large-scale assessments, computer-based testing will continue to increase in use. Great strides have been made in how technology can enhance the testing experience, especially with the amounts of money that large-scale assessment companies have to continue the research and development of new additions of technology (Beller, 2013). What started as a means to increase simple efficiency has made way for much greater complexity in the use of technology for large-

scale assessments, and this trajectory can be appealing to the users of classroom-based assessments.

### **Technology and Classroom Assessments**

There is a clear distinction between large-scale and classroom assessments. As opposed to the large-scale assessment purpose of accountability and comparison, the foundation for classrooms assessments is to determine what has been learned, and adjust the learning environment and instruction in response to information gathered from the classroom assessment (Quellmalz & Pellegrino, 2009). This is also known as formative assessment. In formative assessment, changes in instruction and conditions of learning are based on frequent classroom assessments, which have been shown to have a significant positive effect on student achievement (Black, Harrison, Lee, Marshall, & Wiliam, 2004). Given the number of assessment opportunities that a classroom teacher could potentially come across during one day, multiplied by the number of students in his or her classroom, it is easy to see how administering, collecting, scoring, providing feedback, and then adjusting instruction using PBTs would be a daunting task (Brown, Hinze, & Pellegrino, 2008).

When compared to the trajectory of technological development for large-scale assessments, many classroom assessments are still in the initial phase of focusing on CBTs for the sake of efficiency. With CBTs in classroom assessment, opportunities for streamlined data collection, immediate scoring, and quick feedback to both the student and the teacher are now easily accessible (Brown, Hinze, & Pellegrino, 2008; Ripley, 2009). Learning management systems (LMS), such as Moodle or Blackboard, support teachers in using technology to share information and resources, including administration

of CBT polls, quizzes, and tests. These assessments are instantly graded within the LMS and the teacher and student receive immediate results (Halvorson & Shapiro, 2012). Student response systems (SRS) are another option to support classroom CBTs. From sets of classroom “clickers,” or handheld polling devices, to fully interactive systems using tablets and smartphones, SRS allows for immediate collection, scoring, and feedback not possible with PBTs (Waters, 2012). Teachers can also create their own CBTs. Google Forms lets teachers create tests, quizzes, and questionnaires, and the web-based app automatically scores, provides feedback, and can even analyze trends, giving teachers the timely information they need to make instructional decisions (Waters, 2012).

With all of these options to support CBT in the classroom, it is still up to the teacher to determine the content of the assessment. Quick administration, scoring, and feedback are useless if the assessment is not valid and reliable in determining student performance. In order to assist in making appropriate instructional decisions, teachers have been implementing a tool called curriculum-based measurement.

### **Curriculum-Based Measurement**

Curriculum Based Measurement (CBM) is a standardized process designed to evaluate student achievement in basic academic skills (Deno, 2003). CBM was developed in the mid-1970s by the University of Minnesota’s Institute for Research on Learning Disabilities (IRLD) as a way to bridge the gap between instruction and assessment, providing special educators with a quick and efficient way to monitor student progress in basic skills (Deno, 1985). CBM is now used by both general education and special education teachers and can be used for screening with all students to identify which might be at risk academically, predict performance on high-stakes assessments, and evaluate the

effectiveness of interventions (Christ & Vining, 2006; Deno, 2003). When used as a screening tool, teachers administer CBM measures in the fall, winter, and spring for all students, using the information provided to determine if students are on track to meet grade level academic goals. If it appears that a student is not on track to meet grade level academic goals, then CBM can be used as a progress monitoring tool on a more frequent routine, from once a month to once or twice a week. This may also lead to the implementation of an intervention, with the CBM monitoring progress and determining the effectiveness of the selected intervention (Lembke & Stecker, 2007). CBM has proven to be a reliable measure that is sensitive to student growth, allowing for teachers to make appropriate instructional decisions in a timely manner (Shapiro, 2004). CBM is intended to be a reflection of the curriculum, and can be used in the areas of reading, writing, and mathematics.

Mathematics CBM (M-CBM) can be used with pre-K to high school students, starting with early numeracy measures and computation, through concepts and applications and algebra measures (Clarke & Shinn, 2004; Foegen, Olson, & Impeccoven-Lind, 2008; Lembke & Foegen, 2009). Early numeracy CBM measures include student demonstration of oral counting, number identification, quantity discrimination, and missing number (Lembke, Foegen, Whittaker, & Hampton, 2008). These early numeracy skills have been identified as indicators of number sense, and are essential to the progression of mathematical proficiency (National Mathematics Advisory Panel, 2008). Computation measures consist of single skill or mixed basic facts or multi-step problems of addition, subtraction, multiplication, and division (Lembke & Stecker, 2007). These measures of arithmetic fluency and proficiency show both accuracy and the use of

efficient counting strategies, and students who lack these skills often display learning difficulties in mathematics (Gersten, Jordan, & Flojo, 2005). M-CBM in concepts and applications require students to successfully apply mathematical knowledge to problems of applied computation, charts and graphs, measurement, money, time, quantity, and word problems (Lembke & Stecker, 2007). In order to be successful on concepts and applications measures, one must have knowledge of the concepts, strategies, and facts, and be able to pull them together in order to solve problems (Thurber, Shinn, & Smolkowski, 2002).

Administration of M-CBM measures ranges from 1 to 8 minutes. Individual administration is required for early numeracy measures, but M-CBM measures for older students are generally group administered (Lembke, Hampton, & Beyers, 2012). M-CBM can be scored by counting correct digits, correct problems, or correct filled-in blanks. Students are encouraged to complete problems in the order of presentation, but can skip those that are too difficult (Lembke & Stecker, 2007).

M-CBM measures can be defined as subskill mastery measures or general outcome measures (Christ & Vining, 2006). Subskill mastery measures assess single skills, such as single digit multiplication facts. With the narrow focus of subskill mastery measures, students are expected to make quick progress in a short span of time, and can be useful when it has been determined that a student has specific skill deficits (Christ & Vining, 2006). As discussed earlier, boosting subskills is necessary for success on computation measures, which could lessen difficulties in mathematics (Gersten, Jordan, & Flojo, 2005). General outcome measures assess a broad range of skills that are expected to develop over time and are reflective of multiple skills gained over an

academic year (Christ, Scullin, Tolbize, & Jiban, 2008). With the broad range of skills included in a mathematics general outcome measure, the rate of progress is not as quick as to be expected with subskill mastery measures, but are much more indicative of progress within the curriculum (Christ & Vining, 2006).

The types of problems used in M-CBM are developed through two approaches: curriculum sampling or robust indicators (Foegen, Jiban, & Deno, 2007). Curriculum sampling uses problems and applied skills that are expected in each grade. This allows for a direct link to the curriculum so teachers receive immediate feedback and can design instruction to teach specific skills. Robust indicators are made up of skills that represent general markers of proficiency in mathematics, instead of directly linking to the curriculum (Christ et al., 2008). Use of robust indicators can be beneficial, as they can be used over a span of grade levels and are not connected to any specific curriculum (Foegen, Jiban, & Deno, 2007).

Early research on M-CBM examined basic computation in addition, subtraction, multiplication, and division, with results showing moderate correlations with elementary and middle school standardized achievement tests in mathematics (Foegen & Deno, 2001; Thurber, Shinn, & Smolkowski, 2002). Even with moderate correlations, the use of basic facts as a measure of overall mathematics achievement has not been as widely accepted, most likely due to the reforms in mathematics education that have removed the focus from fact fluency (Jiban & Deno, 2007). Many popular approaches to teaching mathematics focus on conceptual knowledge and problem solving, even with basic math facts (Woodward, 2006).

Even though the use of basic facts as a measure of overall mathematics has not been held in the highest regard, research has shown that fluency with basic facts is an important skill. There are advantages for students who display mathematical fluency, mainly the ability to understand and complete complex mathematical tasks. More complex mathematical tasks are easier for students with strong basic math skills. Basic skills are necessary as a foundation for more difficult mathematics skills (Fuchs et al., 2006; Vukovic & Seigel, 2010). Lack of fluency indicates inefficient counting strategies. If students have to count on their fingers or draw pictures in order to solve basic facts, they will have difficulty understanding more complex skills (Bryant, Hartman, & Kim, 2003; Gersten & Chard, 1999). The Common Core State Standards for Mathematics also includes standards for fluency. By the end of third grade, students are expected to fluently add and subtract with numbers 1-20, and have committed all sums of two one-digit numbers to memory. By the end of third grade, students are expected to be fluent in multiplication and division within 100, and have memorized all products of two one-digit numbers (Common Core State Standards Initiative, CCSSI, 2010).

### **Technology and CBM**

CBMs were designed to be quick instruments with standardized administration and scoring procedures but issues still arose with time commitment, consistent administration, and data collection (Fuchs, Fuchs, & Hamlett, 1994). They are quick measures, when referring to student involvement and time, but there is still quite a bit of teacher time involved. Significant time is required to organize and administer assessments, from preparing materials to scoring and graphing data. It can also be difficult to analyze the data and know how it translates into instructional changes (Fuchs,

1998). Computer applications were designed to help alleviate implementation problems by automating data collection, graphing, and data analysis. In the early stages of use, studies revealed that while teachers were pleased with the use of computer-based CBM, it was actually less efficient than paper-based CBM due to issues in the graphing procedures (Fuchs, Fuchs, Hamlett, & Hasselbring, 1987). The next step in technology and CBM included generation, administration, and scoring of CBM probes, with the intention of helping teachers spend less time administering and scoring measures, and more time making instructional decisions (Ferguson & Fuchs, 1991). In order to assist in making instructional decisions about what to teach, computer-based skills analysis was added. Using this analysis, teachers made better instructional plans, leading to increased student achievement. This increase in achievement was seen in both reading and mathematics (Fuchs, Fuchs, & Hamlett, 1989; Fuchs & Fuchs, 1990). Fuchs and Fuchs (2001) then developed expert systems to go along with computer-based CBM. Expert systems are designed to assist the teacher in deciding not only what to teach, but how to teach. Since then, the computerization of CBM probes has continued, leading to the development of many computer-based CBM tools and supports (Goo, Watt, Park, & Hosp, 2013).

AIMSweb, mCLASS, easy CBM, Edcheckup, FastBridge and are available options for computer-supported M-CBM tools (Goo et al., 2013; Phillips, Shinn, & Ditkowsky, 2014; Ysseldyke & McLeod, 2007). AIMSweb and mCLASS have computer-based supports for scoring, graphing, and data analysis, but do not have options for computer-based administration. EasyCBM and Edcheckup have the option for computer-based administration, with students using the computer to complete



assessments. Both allow students to print and use a paper-based format instead of computer-based. EasyCBM assessments are presented in a multiple choice format. Edcheckup employs a cloze math procedure, where students are expected to fill in the blanks in number sentences. FastBridge contains a fact fluency measuring tool that can be completed using a laptop or tablet.

Many of the existing options for computer-based assessments require some level of purchase, which makes it beyond the financial capabilities of some classrooms, buildings, or school districts. Teachers and school district personnel who are skilled in the use of technology can create their own computer-based M-CBM with minimal to no cost (Ysseldyke & McLeod, 2007). Free web-based tools such as Google Forms allow for teachers to easily create forms that are automatically scored and entered into a spreadsheet. Computer-based options include creating forms in Microsoft Office, which can also be automatically scored. However, due to the lack of research and opportunities to make comparisons based on scores obtained through different modes, these options emphasize the importance of examining for mode effects.

### **Mode Effects and Assessment**

Identical computer-based and paper-based assessments should result in the same scores. If the scores are not the same, or at least close to the same, this difference in scores is a test mode effect. (Clariana & Wallace, 2002). Something other than the construct intended to be measured is factoring into the result due to the mode of the assessment. A test mode effect is a specific form of construct-irrelevant variance (Huff & Serici, 2001). Research on mode effects of assessment involves examining why the difference in scores occurs. Existing research can be organized in two categories:

research on technology issues related to mode effects and research on participant issues related to mode effects. Technology issues refer to concerns due to the interface of CBT. Examples of technology issues include, but are not limited to, font size or color, screen resolution, or the ability to skip questions and change answers. Examples of participant issues include, but are not limited to, familiarity with technology, socioeconomic status, or age. It is important to note that although the research on mode effects spans back decades and over a variety of types of assessments, the studies reviewed in depth in the next few sections were limited to those published after 2000 in order to include studies that most reflect current technological practices. There are also studies that examine mode effects of assessment for different purposes, including surveys, psychological tests, and work-based assessments. To fit the purposes of this study, only studies that included academic assessments were included.

### **Technology Issues**

Many studies have looked at how technological factors may relate to mode effects. Much of the research in the 1970s through the 1990s focused on screen resolution and font characteristics. Gould, Alfaro, Finn, Haupt, and Minuto (1987) found that low resolution screens resulted in lower reading performance and higher levels of eye fatigue when compared to PBT. Font characteristics were also important to compare, because the most commonly used fonts on the computer (Times New Roman and Arial) were originally designed for high resolution on paper. However, on screens with lower resolution, these fonts become less legible and reading performance suffers (Bernard & Mills, 2000). With rapid development and improvement in technology related to computer displays, mode effects related to screen resolution are generally no longer an

issue (Leeson, 2006); therefore, those studies will not be included in this literature review.

One technological issue that has continued to show up in the research comparing PBT and CBT has to do with item review. Item review is the ability to review, skip, and change answers on CBT versions of assessments. On PBTs, there is no way to control how individuals move through the test. Skipping questions, reviewing responses, and changing answers is simple. On CBTs, depending on how the testing software is configured, it can be impossible to skip questions or to go back and review or change an answer once it has been submitted. Vispoel (2000) maintained that the ability to review and change CBT test items increases test score validity because test-takers are able to correct for typing errors, items that may have been misunderstood, lapses in memory, and reconceptualization of answers.

Previous studies found that the ability to review and change items on CBTs did not have an effect on performance, but did have an effect on speed (Eaves & Smith, 1986; Luecht, Hadadi, Swanson, & Case, 1998; Spray, Ackerman, Reckase, & Carlson, 1989). A study examining comparability of PBT and CBT unit tests in an undergraduate psychology course found that CBT tests did not have an effect on performance. A PBT version was given, along with three versions of CBT: presenting all questions at one time (much like paper), presenting a single question with the ability to review, and presenting a single question with no option to review. There was no difference in score but a significant difference in completion time, with the single question and no review option as the fastest of all CBT formats. However, all of the CBT formats resulted in longer completion time when compared to the PBT version (Bodmann & Robinson, 2004),

indicating that the switch from PBT to CBT has an impact on speed, whether allowed to review and change or not.

In a study of PBT and CBT formats of the Graduate Record Exams (GRE), Goldberg and Pedulla (2002) compared a PBT format of the GRE to two CBT formats. One CBT format allowed the participant to review and make changes, but the second CBT format did not allow review or changes. On the PBT version of the Verbal subtest, 96% of the participants reached the last question. On the CBT version with the ability to review and change answers, 60% of the participants reached the last question, and on the CBT version with no review, 78% reached the last question. This pattern also held for other subtests. In terms of raw scores, on all subtests, scores obtained through PBT administration produced significantly higher than those obtained through CBT with or without review (Goldberg & Pedulla, 2002).

There is an important difference to note in these two studies and the difference in findings, both relating to differences in speed based on the mode of assessment. The first study allowed participants to complete the entire test, with no time limits, but kept track of the amount of time required to complete the test. This led to a significant difference in the time required, but not a significant difference in performance on the test. We cannot infer what the comparisons in performance might have been if every participant was given the same amount of time. The second study provided a specific amount of time for participants to answer as many questions as possible, but when the time had expired, no more questions could be answered. This study found a significant difference in performance based on raw scores, but we do not know how this might have changed if participants were allowed to answer all questions without a time limit.

## **Participant Factors**

Participant factors in the research on mode effects include socioeconomic status, parent education level, sex, age, culture, race, familiarity with technology, and ability levels in area being tested, with most studies including research on multiple participant issues. For a breakdown of each study and the factors examined, see Table 1.

Using eligibility for free and reduced lunch (FRL) as a proxy for lower socioeconomic status, Pomplun and Custer (2005) studied the comparability of scores from PBT and CBT versions of a K-3 multiple-choice reading test (Basic Early Assessment of Reading). Overall, participants scored significantly higher on the PBT. At each grade (K-3), students who were eligible for FRL scored higher on PBTs and had larger score differences than their non-FRL peers, but the differences decreased as grade level increased. The researchers stated that this decrease in differences as grade level increased may be due to increased amounts of practice and familiarity with the computer (Pomplun & Custer, 2005), although no measure other than FRL status was used to draw this conclusion.

A study in Australia compared performance of Aboriginal and non-Aboriginal elementary students on PBT and CBT versions of assessments meant to measure the level of proficiency with basic math facts. These two versions were not identical assessments. The PBT was the Operations subtest of the Key Maths Test. To reflect the skills represented in the PBT, the researchers created the CBT version. Neither assessment was timed, but the CBT did record the time required to complete all problems. The PBT was given first, with non-Aboriginal students significantly outscoring Aboriginal students. For each subtest, the effect sizes were very high: addition ( $d = 1.01$ ), subtraction, (4.22),

multiplication (2.87), and division (3.32). Participants were then given the CBT test and also given at least four opportunities to complete each of the CBT subtests in addition, subtraction, multiplication, and division, with the idea that when given multiple opportunities on the CBTs, the resulting scores would be more accurate than when given only one opportunity. At the end of the repeated opportunities, Aboriginal students outperformed non-Aboriginal students on the CBT version. The Aboriginal participants also showed significant increases in mean score from the first CBT opportunity to all subsequent sessions. Results showed that Aboriginal students performed better than non-Aboriginal peers, with high effect sizes: addition ( $d = 0.71$ ), subtraction (2.58), multiplication (1.00), and division (1.09). Authors believe that results from this study imply that the CBT may be less culturally biased than the PBT (Hippisley, Douglas, & Houghton, 2005). However, this interpretation would be inaccurate. If one were to decide that the CBT might be less biased for Aboriginal students, then it would also be true to say that the PBT is less biased for non-Aboriginal students.

In a study with 10-year-olds in England, researchers examined the comparability of PBT and CBT formats with a mathematics assessment composed of multiple-choice and short answer questions (Hargreaves, Shorrocks-Taylor, Swinnerton, Tain, & Threlfall 2004). Identical tests were created in PBT and CBT formats. Participants also completed a questionnaire about their familiarity with computers. The study found scores were slightly, but not significantly higher on CBT versions of a mathematics assessment, but could find no patterns to connect computer skill and performance. However, in the within-subjects analysis, there were a small number of children (24 out of 260) who obviously performed better on one version over another, suggesting that even though the

overall results reflect no significant difference, the change in formats does affect some students. Authors also noted that all participants completed the PBT version first, so there may have been a practice effect.

Studies have also been completed to examine the comparability of PBT and CBT versions of the National Assessment of Educational Progress (NAEP). Bennett et al. (2008) studied PBT and CBT versions of the mathematics section of the NAEP. Overall, there were significant differences in PBT and CBT scores, with greater variability of CBT scores. Comparisons were made by sex, race/ethnicity, parent education, region of the country, type of school, and computer proficiency. For the entire sample, scores were higher on PBT. The difference was statistically significant, but small in terms of effect size (0.14). When examining by subgroups, there were no significant differences between PBT and CBT, except for those who reported having at least one college graduate parent. For this group, scores were higher on the PBT version, but the effect size (0.21) was still considered to be small (Bennett et al., 2008).

Horkay, Bennett, Allen, Kaplan, and Yan (2006) also studied PBT and CBT versions of the NAEP, but for the writing section. There were no overall differences in PBT and CBT scores, and no differences in score by sex, race/ethnicity, parent education, region of the country, type of school, or computer proficiency. Computer proficiency was determined using a measure of word processing skill. Participants who took the CBT produced slightly longer essays, but only about nine words longer. Students from “urban fringe/large town” areas performed significantly higher on the PBT, but the effect size (0.15) was very small.

Comparability of PBT and CBT results on the mathematics section of the Kansas statewide assessment has also been examined. The PBT and CBT were both multiple-choice assessments. There were no overall differences when comparing performance on PBT and CBT. There were also no differences in performance associated with sex, academic placement (general education, gifted education, special education), or socioeconomic status. Differential item functioning analysis did show that there were some items that were observed to be more difficult in the CBT mode, but did not have an impact on overall test scores. The authors stated that a possible reason for this difference was the size of the items, as they were all large enough to require scrolling to view the entire item (Poggio, Glasnapp, Yang, & Poggio, 2005).

In order to compare End-of-Course assessments in Algebra I and English I, Lottridge, Nicewander, and Mitzel (2011) had participants complete PBT and CBT versions of a multiple-choice assessment created by the state. Computer skills, reading scores, math scores, areas of the state, free and reduced lunch status, special education/gifted status, and ethnicity were all included in the analysis. The within-subject results suggest that the PBT and CBT versions are measuring the same content with the same levels of reliability, but the CBT was shown to be slightly more difficult than the PBT, although not significantly so. The authors felt that even though there may be some construct irrelevant variance connected to the CBT, these differences could be removed through a process of equating scores.

In an examination of comparability of PBT and CBT formats of the GRE, Goldberg and Pedulla (2002), found that scores on PBT formats were higher than scores on CBT formats. Participants with higher levels of computer familiarity outperformed



those with lower levels of computer familiarity on CBT versions. For this study, computer familiarity was determined through a series of survey questions regarding participant familiarity with various computer skills.

In a study with participants from an undergraduate computer fundamentals course, participants completed a 100-item multiple-choice test was presented in PBT and CBT formats. Results indicated performance on CBT was significantly higher than PBT. A further examination of participant characteristics found no difference based on sex or computer familiarity, but found a significant difference based on content familiarity. High performing students were helped by the CBT version, while there was no significant difference for low performing students when comparing performance on PBT and CBT tests (Clariana & Wallace, 2002). Even though content familiarity was found to be a possible contributing factor for this study, it is important to note that the content of this course is based on computer fundamentals. In a similar study conducted in an undergraduate psychology course, Mason, Patry, and Bernstein (2001), found no difference in performance between PBT and CBT versions of unit tests. They used a computer attitude inventory to examine the effects of familiarity with computers, but all participants involved felt confident in their computer skills. This similarity in performance and confidence in computer skills may be a more accurate reflection of the similarity of students participating in the study instead of showing that computer skills may be a factor in differences between PBT and CBT scores.

A few patterns start to emerge in the reviewed studies. In most instances, research on mode-effects has been completed using a multiple-choice format. Of the 11 studies reviewed in this section, seven were constructed of multiple-choice questions only. The

CBT formats also seem to consistently decrease speed when compared to PBT formats. There are still many inconsistencies between studies. One study found that the CBT format resulted in higher performance, four studies found that PBT formats resulted in higher performance, and six studies found no significant difference between formats. These inconsistencies lead to the need for further studies in the area of mode effects of assessments. The absence of studies on CBM and mode effects of assessment also suggest a need for research in that area.

**Table 1***Summary of Studies on Mode Effects of Assessment*

Article	Sample	Assessment	Results
Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008	N = 1900; 8 <sup>th</sup> grade	PBT and CBT block of 26 multiple-choice, short and long constructed response NAEP Math questions; background questionnaire	CBT group scores were significantly more variable than PBT; students with one college graduate parent performed significantly higher than other groups on PBT
Bodmann & Robinson, 2004	N = 58; undergraduates in psychology course	All CBT multiple choice; interface 1- presents all questions at once, 2- single question, can review, 3- single question, no review	No mode effect based on score; mode effect for completion time; single question with no review faster
Clariana & Wallace, 2002	n = 105; undergraduates in computer fundamentals course	100 item multiple choice; CBT answers reviewable; PBT 6-7 questions per page; Distance Learning Profile	Significant difference in scores with CBT scores higher than PBT; no mode effect for sex, computer familiarity, competitiveness; mode effect for content familiarity
Goldberg & Pedulla, 2002	N = 222; undergraduate volunteers	3 versions of the GRE- PBT, CBT can review and change, CBT no review; Computer familiarity scale	Participants in PBT group outperformed both CBT groups; CBT had significant effect on speed (slower than PBT); higher computer familiarity outperformed lower computer familiarity
Hargreaves, Shorrock-Taylor, Swinnerton, Tait, & Threlfall, 2004	N = 260; 10-year-olds	Four mathematics assessments, 2 CBT and 2 PBT, mix of multiple choice and short answer; Computer skill questionnaire	Scores were higher on CBT, but not statistically significant

Hippisley, Douglas, & Houghton, 2005	N = 290; Australian elementary students	Basic arithmetic skills in PBT and CBT format; constructed response	Non-Aboriginal participants scored higher than Aboriginal participants on PBT, but Aboriginal participants scored higher than non-Aboriginal participants on CBT.
Horkay, Bennett, Allen, Kaplan, & Yan, 2006	N = 1313; 8 <sup>th</sup> grade students	NAEP essay writing measure, PBT and CBT, 25 min for each measure; online computer skills measure	No score effect for delivery mode; no score differences for sex, race/ethnicity, parent education level, school location, FRL, school type; slight increase in length on CBT
Lottridge, Nicewander, & Mitzel, 2011	N = 2101 (Algebra I) N = 1527 (English I); 8 <sup>th</sup> and 9 <sup>th</sup> grade students	State End of Course assessments in Algebra I and English I, CBT and PBT versions; all multiple choice	For both Algebra I and English I, scores were slightly higher on PBT, but no significance; follow up using PSM found similar results
Mason, Patry, & Bernstein, 2001	N = 27; undergraduates in an Introductory Psychology course	Multiple choice unit tests; Computer attitude inventories	No significant difference in scores or computer attitude
Poggio, Glasnapp, Yang, & Poggio, 2005	N = 644; 7 <sup>th</sup> grade students	Kansas statewide mathematics assessment, PBT and CBT versions; multiple choice	Very little difference in performance on PBT or CBT; no differences based on academic placement, SES, sex; some differences with specific items in CBT
Pomplun & Custer, 2005	N = 1959; K-3 <sup>rd</sup> grade	Reading screening tests; multiple choice; CBT items shown one at a time, skipped items readministered	At all grades, students scored significantly higher on PBT; eligibility for free/reduced lunch had larger score differences

## **Issues in Mode Assessment Research**

Kolen and Brennan (1995) suggested that for every assessment available as PBT and CBT, research on mode effects is necessary. For each assessment presented in both PBT and CBT formats, there will be different factors that contribute to mode effects. The factors contributing to mode effects of multiple-tests may be different from those in an essay or short-answer test. Tests in different content areas can be affected by different factors. Differences in timed or untimed tests may produce diverse results. Couple these differences with the lack of consistent findings in the existing research lead to the need for research before making assumptions regarding comparability. This creates an extra responsibility for test developers and test users (Wang, Jiao, Young, Brooks, & Olson, 2008).

Conducting research comparing modes of assessment can be tricky. Many of the studies cited in this literature review make use of surveys and questionnaires to collect information on participant use and familiarity of technology. One issue in the use of surveys and questionnaires comes from the difficulty in defining technology and relying on a consistent participant understanding of that definition (Bebell, O'Dwyer, Russell, & Hoffmann, 2010). With technology use ranging from basics like word processing to more advanced applications like simulations, self-report of technology use via surveys can be misleading. For example, in 1992 a survey study conducted by the International Association for the Evaluation of Educational Achievement defined a computer-using teacher as one who sometimes used technology with students (Office of Technology Assessment, 1995). This survey study estimated 75% of teachers as computer-using teachers. (OTA, 1995). In 1994, Becker used a survey to collect data on computer-using

teachers, but defined a computer-using teacher as one who has at least 90% of the students using technology. This survey placed 25% of teachers as computer-users (Becker, 1994). The exponential increase in the complexity of technology in recent years makes it even more difficult to accurately measure computer use (Bebell et al., 2010).

Another issue in existing research on the comparability of PBT and CBT formats is in the decision making process employed in making the comparisons. The International Testing Commission (2006) has provided extensive guidelines on the use of CBTs, including guidelines for test publishers, test developers, and test users, in the areas of hardware and software requirements, item presentation, accessibility, scoring and interpreting, and control over test conditions. In regards to comparability of PBT and CBT formats, the ITC provides four guidelines. The two versions must: (1) have comparable reliabilities, (2) correlate between formats, (3) correlate similarly with an external criterion measure, and (4) generate comparable means and dispersions of scores, or be appropriately rescaled to provide comparable scores (International Testing Commission, 2006). Of the eleven reviewed studies, all used some combination of comparison of means and score dispersions to determine comparability, four examined reliabilities, and only one examined correlations with a criterion measure.

### **Overview and Research Questions**

The overall purpose of this study is to examine some of these components presented in the existing research, but expand to participants and content areas that have not been included in the research to date, as well as adhere to all of the guidelines put forth by the International Testing Commission on examining comparability of PBT and CBT formats. The first study objective was to determine if there was a mode effect

between PBT and CBT M-CBM probes. Second, the study examined possible factors that could contribute those mode effects, including age, sex, and overall ability in mathematics.

The following research questions were addressed:

- 1) Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer?
- 2) Are there differential mode effects based on sex, grade level, or ability level?

## **CHAPTER THREE**

### **METHOD**

#### **Overview**

The purpose of this study was to examine the differences in student performance when given identical paper-based and computer-based M-CBM probes and to investigate the differences that might be accounted for by the factors of sex, age, and overall ability in mathematics.

#### **Participants and Settings**

Participants were enrolled in grades 4-5 in two elementary schools in a suburban district in the Midwest. A total of 155 students were enrolled in seven participating classrooms, 76 in fourth grade and 79 in fifth grade. In school one, there were two fourth grade classrooms and two fifth grade classrooms. School two participant classrooms included one fourth grade classroom and two fifth grade classrooms. See Table 2 for demographic information.



**Table 2***Participant Demographics*

<b>Descriptors</b>	<b>Number of Students (%) (<i>n</i> = 155)</b>	<b>Number of Students (%) at School One (<i>n</i> = 93)</b>	<b>Number of Students (%) at School Two (<i>n</i> = 62)</b>
<b>Sex:</b>			
Male	85 (55)	50 (54)	35 (56)
Female	70 (45)	43 (46)	27 (44)
<b>Social Economic Status:</b>			
Students receiving free/reduced lunch	45 (29)	22 (24)	23 (37)
<b>Race:</b>			
Asian/Pacific Islander	8 (5)	5 (5)	3 (5)
Black	10 (6)	6 (6)	4 (6)
Hispanic	15 (10)	6 (6)	9 (15)
White	116 (75)	72 (77)	44 (71)
Two or More Races	6 (4)	4 (4)	2 (3)
<b>Support:</b>			
Special Education	19 (12)	11 (12)	7 (11)
English Language Learner	19 (12)	9 (10)	10 (16)

*Note.* Percentages rounded to the nearest whole number.

Due to absenteeism, 13 students were not included in the final analysis. The final analysis reflects students who participated in all assessments, and included 69 fourth grade students and 73 fifth grade students. There were some analyses that only required certain assessments to be completed, and included 71 fourth grade students and 74 fifth grade students. No consent forms were required, as the University Institutional Review Board determined that all of the materials used in the study were part of regular classroom activities and no identifying information was collected.

**Study Design**

The purpose of this study was to investigate the differences between paper-based and computer-based administration of M-CBM probes and then to examine how sex, grade level, and mathematical ability may contribute to score differences. This required a

mixed-factorial design with both within- and between-subjects variables. Individual performance on PBT and CBT M-CBM was a within-subjects variable. Each participant completed all paper-based and computer-based measures. To control for practice effects, the order of administration was counterbalanced, with three classrooms completing the paper-based M-CBM probes first and four classrooms completing the computer-based M-CBM probes first. Between-subjects variables included sex, grade level, and mathematical ability, as determined through an assessment of math skill.

### **Instruments**

**Group Math Assessment and Diagnostic Evaluation.** The Group Math Assessment and Diagnostic Evaluation (GMADE) is a standards-based, norm-referenced assessment of math skill, designed for use with kindergarten to high school students. The GMADE represents process standards, including representation, connections, problem solving, communication, and reasoning and proof, as well as the content standards of numbers and operations, algebra, geometry, measurement, and data analysis and probability. The entire battery is made up of three subtests: Concepts and Communication, Operations and Computation, and Process and Application. Concepts and Communications assesses proficiency in language, vocabulary, and representations of mathematics. Operations and Computation measures the basic operations of addition, subtraction, multiplication, and division at a level that is appropriate for the grade. Process and Application requires that students understand mathematical language and apply appropriate operations to solve problems. The assessment can be group administered. It is a test of power, not speed, meaning it is untimed, but usually takes 45-90 minutes. The GMADE has been shown to have an alternate form reliability of .89,

test-retest reliability of .90, and an internal reliability of .96 (Center on Response to Intervention, n.d.). The GMADE also has a median criterion related coefficient of .83 and a median predictive validity coefficient of .81 to the Iowa Tests of Basic Skills, TerraNova, and Iowa Tests of Educational Development (Center on Response to Intervention, n.d.).

**CBM Probes.** I created PBT and CBT versions of M-CBM probes, first for a pilot study and then for use in this study. Two forms were produced, each containing 100 multiplication facts with multiplicands ranging from zero to 12, in alignment with district expectations. A random worksheet generator ([www.mathscore.com](http://www.mathscore.com)) provided the order of the problems. Problems were presented in a vertical format, arranged with five problems per row. Two minutes were allowed for each of the two forms.

The PBT versions were comprised of seven rows of problems, with 35 problems on the first two pages and 30 on the third page. The worksheets were created using Microsoft Word Forms. Participants responded by writing the answer in the space given below the problem. See Appendix B for the PBT versions.

The items on the two PBT probes were used to create parallel CBT probes. Problems were displayed in the same order and arrangement as the PBT forms, with five problems per row to a total of 100 problems. Fillable web-based probes were created using web-interface forms. Computer probes were completed by using the Tab key to move to the next problem and the number keys to enter the answers.

The M-CBM probes used in the pilot study have shown evidence of strong reliability, as shown in Table 3 (Hensley, Rankin, & Hosp, in review). Probes were administered during mathematics instructional time, following standardized CBM

administration procedures (Hosp, Hosp, & Howell, 2007) described below and in Appendix A.

**Table 3**

*Reliability of Pilot Study Measures*

Mode	Probe		
	1 and 2	2 and 3	1 and 3
Paper-Based	.89	.88	.85
Computer-Based	.86	.88	.83

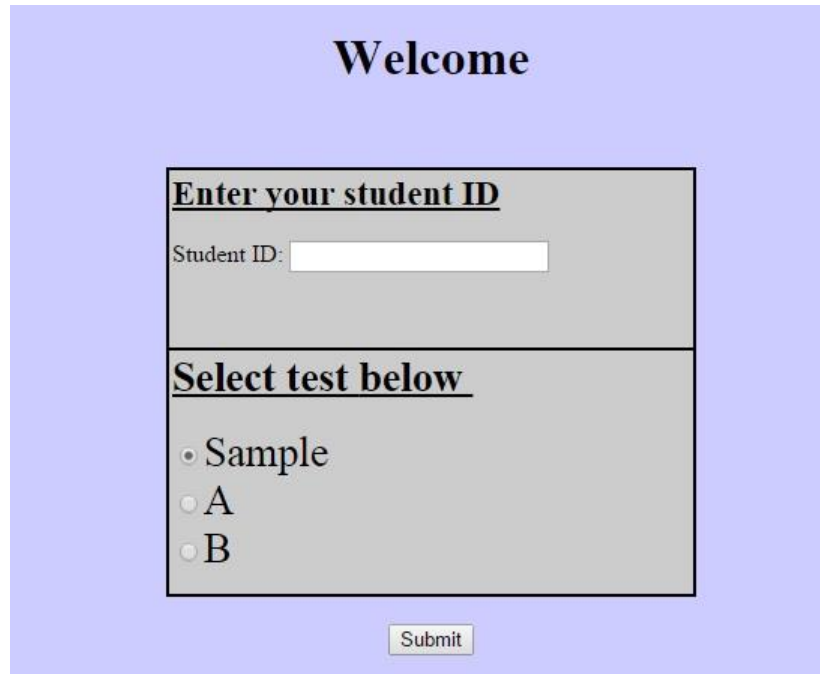
**Procedures**

I trained participating teachers in the administration of PPT and CBT probes using standardized CBM procedures. The standardized directions for the PBT probes included a highlighted script (Appendix A) to read aloud and cues for what the teacher should look for while administering the probe. Two minutes were given for each probe. The order of the probes was counterbalanced between PPT and CBT, as well as the order of probe 1 and probe 2 randomized within the different modes of PPT and CBT. Three classrooms completed the PBT probes first and four classrooms completed the CBT probes first. Both sets of probes were completed within two school days.

Appendix C shows the directions used with CBT probes. Figure 1 shows the welcome screen, where participants enter their student ID and select from a sample test, test form A, or test form B.

## Figure 1

### *Welcome Screen for CBT Probes*



The image shows a welcome screen for CBT probes. It has a light blue background. At the top center, the word "Welcome" is written in a bold, black, serif font. Below this, there is a gray rectangular box with a black border. Inside this box, the text "Enter your student ID" is written in a bold, black, serif font. Below this text is a white text input field with the label "Student ID:" to its left. Below the input field, the text "Select test below" is written in a bold, black, serif font. Below this text are three radio button options: "Sample", "A", and "B". The "Sample" option is selected, indicated by a small black dot inside the radio button. Below the gray box, centered, is a small gray button with the word "Submit" written in a black, sans-serif font.

Participants were instructed on how to complete the CBT probes since this was new to all of them. Through the use of the sample probe, participants were able to learn how to navigate through problems using the Tab key and number keys. The Tab key was marked with a brightly colored sticker. Participants could choose to use the numbers across the top of the keyboard or the 10-key pad. Figure 2 shows the CBT sample test. Participants were instructed on how to submit their completed probes using the Submit button.

**Figure 2**

*CBT Sample Test*

$\begin{array}{r} 8 \\ \times 3 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 5 \\ \times 7 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 7 \\ \times 5 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 4 \\ \times 8 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 6 \\ \times 9 \\ \hline \end{array}$ <input type="text"/>
$\begin{array}{r} 4 \\ \times 2 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 1 \\ \times 1 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 6 \\ \times 3 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 0 \\ \times 4 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 8 \\ \times 5 \\ \hline \end{array}$ <input type="text"/>
$\begin{array}{r} 12 \\ \times 6 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 11 \\ \times 8 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 6 \\ \times 7 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 5 \\ \times 11 \\ \hline \end{array}$ <input type="text"/>	$\begin{array}{r} 4 \\ \times 12 \\ \hline \end{array}$ <input type="text"/>

After completing both PBT and CBT probes, all participants completed the G-MADE subtests within one week. Teachers administered the G-MADE subtests using the scripted directions provided within the Teacher's Manual for the G-MADE.

PBT and CBT probes were scored by counting problems correct. I decided to use problems correct instead of digits correct because in examining fluency of basic facts, the problem is either correct or incorrect. The pilot study found that the results calculated using digits correct mirrored those using problems correct (Hensley, Rankin, & Hosp, in review), so with the emphasis on fluency of basic facts, the decision was made examine only problems correct. PBT probes were scored by the author and a retired educator with experience in scoring M-CBM probes. Both research team members scored all probes, and interscorer agreement was calculated at 97%. Disagreements were reconciled to

create the final dataset. CBT probes were scored automatically. The two scores from each set of measures were averaged to create a mean score for each set, providing a more stable estimate of performance. Scores were also rank-ordered based on PBT, since that is the traditional method of M-CBM administration, and those rank-orders were compared to CBT rank-orders.

### **Analysis of Results**

Descriptive statistics were calculated, including means and standard deviations for PPT, CBT, and G-MADE measures. To answer research question one, the statistics were analyzed according to the guidelines set forth by the International Testing Commission (2006) regarding the examination of comparability of PBT and CBT modes of assessment. This included an examination of (1) reliabilities, (2) correlations between modes, (3) correlations with external criterion measures, and (4) comparable means and dispersion of scores. Parallel-forms reliability was calculated to determine reliability within the same mode. Pearson product-moment correlations were conducted to ascertain correlation between modes of assessment and between each mode and the GMADE, which was used as the external criterion measure. In comparing the correlations between modes and the GMADE, Meng's Z-test was used to determine if overlapping correlations with the GMADE were significantly different (Meng, Rubin, & Rosenthal, 1992). As one additional measure of correlation, rank-order change between the PBT and CBT versions were compared using the Kendall rank correlation coefficient, allowing for an evaluation of the agreement between two sets of ranks (Kendall & Gibbons, 1990). A dependent samples t-test was used to analyze for comparable mean problems correct for PBT and CBT.

To answer research question two, a mixed design analysis of variance with repeated measures on the second factor (mode) was used to analyze differential mode effects for dichotomous predictors of gender and grade. Quantile regression was used to analyze the continuous variables of math proficiency levels, as determined by the G-MADE, and the correlation with performance on either the PBT or CBT M-CBM. Quantile regression allows for the examination of two variables across the entire distribution of scores, including those at the high and low end of the distribution of scores, instead of an estimation based on the mean of the whole group, meaning that correlations at different quantiles can be examined (Koenker, 2005). The data were organized into 11 quantiles (.01, .1, .2, .3, .4, .5, .6, .7, .8, .9, .99), through the use of Statistical Analysis Software (SAS) Version 9.4. Correlations between performance on the PBT M-CBM and the GMADE and the CBT M-CBM and the GMADE were calculated at each grade level using the Pearson product-moment correlation coefficient. Quantile plots were then created by transferring the data from SAS to Excel to create line graphs.



## **CHAPTER FOUR**

### **RESULTS**

This chapter describes the results of the analysis of the data obtained in a study of 4<sup>th</sup> and 5<sup>th</sup> grade student performance on mathematics assessments given using different modes (PBT and CBT). The purpose of this study was to examine differences in performance based on mode of assessment. The analysis included an investigation of factors that may contribute to differences in performance across modes, including sex, age, and overall ability in mathematics.

#### **Findings for Research Question 1**

Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer?

#### **Comparability Results**

In order to determine if the scores obtained with PBT and CBT test modes were comparable, psychometric properties of the two PBT and two CBT forms were examined. For paper-based and computer-based assessments to be considered equivalent, it is necessary to show that both versions

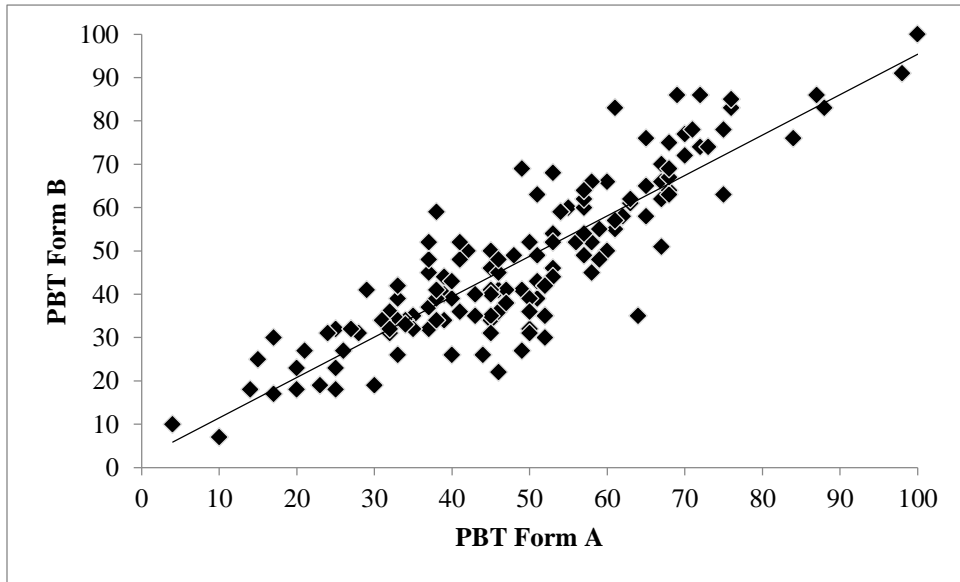
- have comparable reliabilities,
- correlate with each other,
- correlate similarly with an external criterion measure, and
- generate comparable means and score dispersions or can be appropriately adjusted to provide comparable scores (International Testing Commission, 2006).

## Comparable Reliabilities

Each participant took two M-CBM multiplication fact assessments in two modes, for a total of four assessments that measured fluency of multiplication facts. Before any comparisons can be made between different modes of assessment, it is necessary to determine if the scores can be reproduced, also known as reliability. There are different types of reliability, but parallel-forms reliability fits most with the needs of this study. The assessments were analyzed for parallel-forms reliability to determine the consistency between assessments that were not identical, but both created to assess multiplication fact fluency. In order to be considered convincing, parallel-forms reliability of above .80 is required (Center on Response to Intervention, n.d.). The parallel-forms reliability between the two paper-based assessments was  $r = .88$  (see Figure 3) and the parallel-forms reliability between the two computer-based assessments was  $r = .82$  (see Figure 4), indicating adequate reliability, but in using Meng's  $Z$ -test to assess whether the correlations were significantly different (Meng, Rubin, & Rosenthal, 1992), it was determined that the differences in the two correlations were statistically significant (Meng's  $Z = 2.25$ ,  $p = .01$ ).

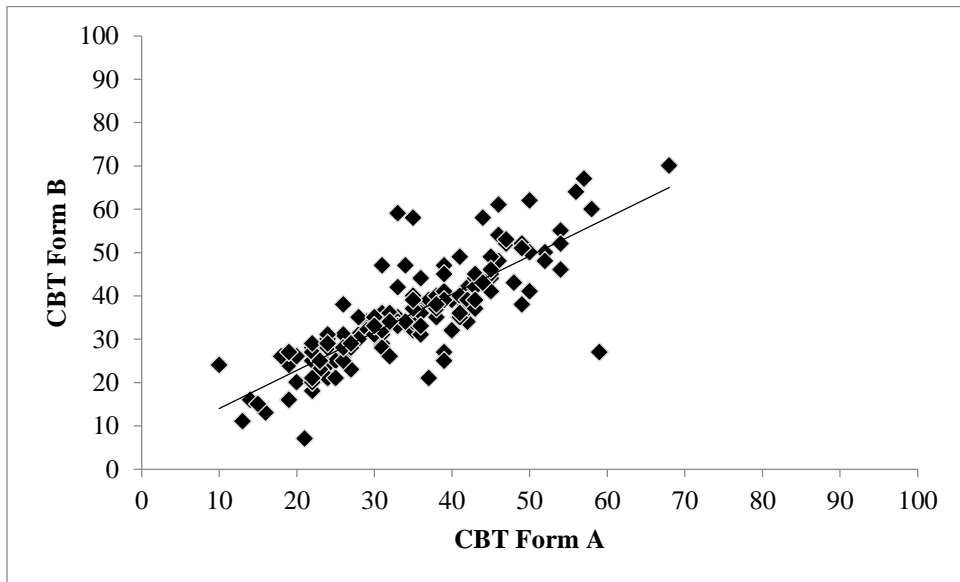
**Figure 3**

*Parallel-Forms Reliability PBT Forms*



**Figure 4**

*Parallel-Forms Reliability CBT Forms*

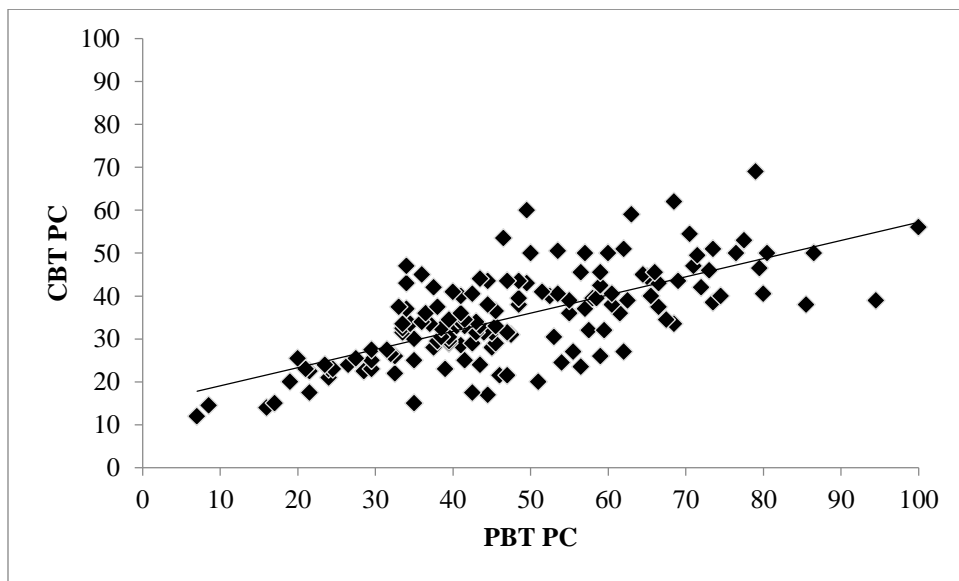


## Correlate Between Test Modes

Pearson product-moment correlations were conducted to ascertain how performance on the PBT mode of assessment correlated with performance on the CBT mode of assessment. A correlation of  $r = .68$  ( $p < .01$ ) was found (see Figure 5 for scatterplot), and while the correlation is statistically significant, it is still a weak association, considering that  $r = .80$  is necessary for evidence of reliability of assessments (Center on Response to Intervention, n.d.).

**Figure 5**

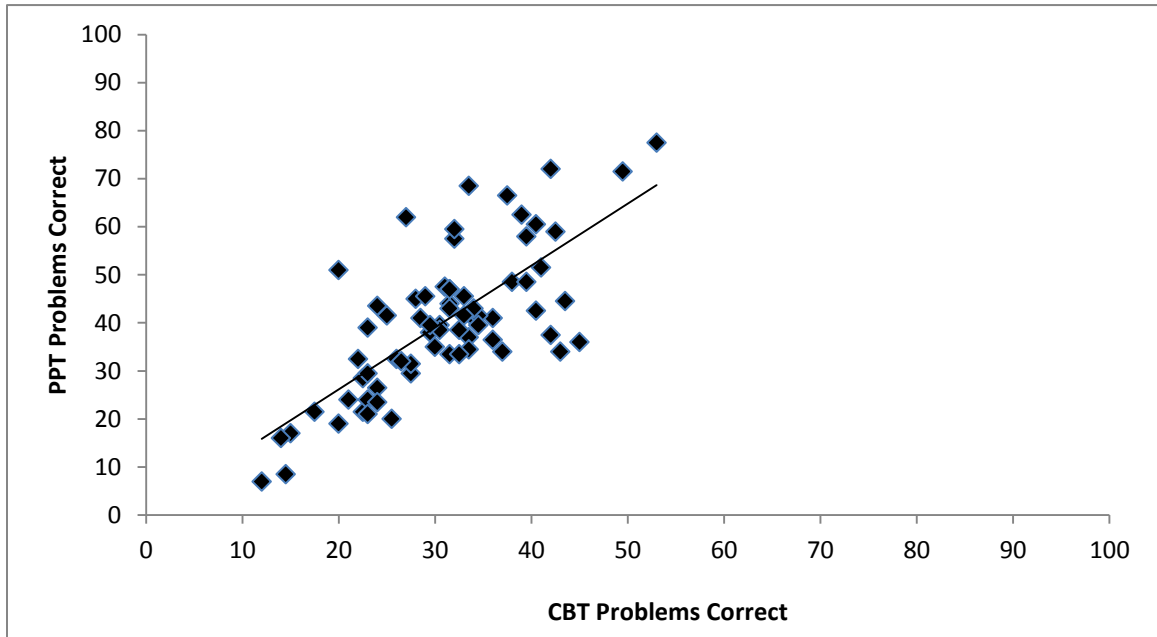
*Scatterplot of PBT and CBT Forms of M-CBM Probes- All Participants*



The correlations for fourth and fifth grade are similar to the overall correlation between PBT and CBT. The correlations are statistically significant, but in terms of assessment reliability, not strong. The correlation between PBT and CBT problems for fourth grade participants was  $r = 0.73$  ( $p < .01$ ), and fifth grade correlation between PBT and CBT was  $r = 0.54$  ( $p < .01$ ). See Figure 6 for fourth grade scatterplots and Figure 7 for fifth grade scatterplots.

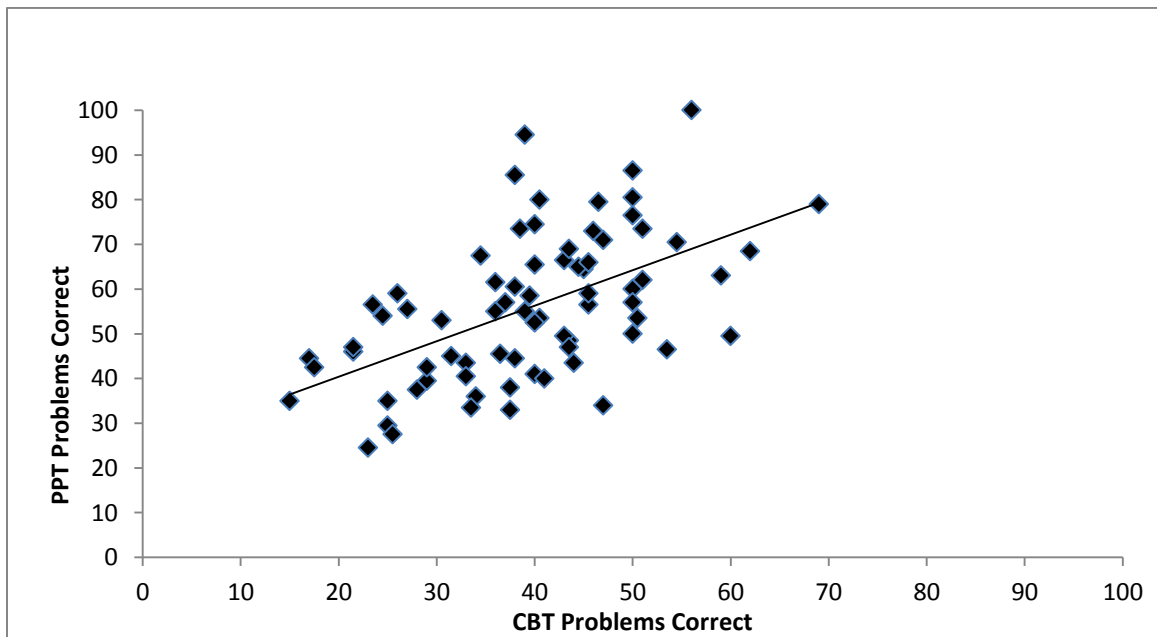
**Figure 6**

*Scatterplot of PBT and CBT Forms of M-CBM Probes- Fourth Grade*



**Figure 7**

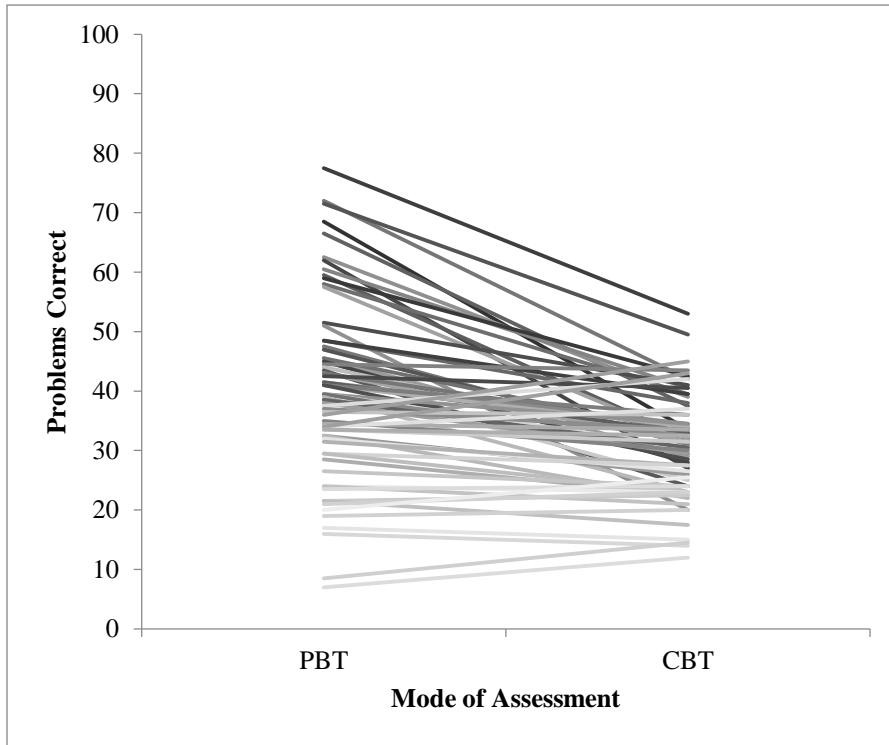
*Scatterplot of PBT and CBT Forms of M-CBM Probes- Fifth Grade*



Another way to look at the correlation of the two assessment formats is through rank ordering. Figure 8 shows fourth grade student rankings on the PBT and on the CBT measures; Figure 9 shows fifth grade students. Both grade levels experienced changes in rank between assessment modes. Appendix D shows scores and ranks for each individual. Just like weak correlations, changes in rank ordering from the PBT to the CBT format would suggest that one format of the test is not measuring the same construct as the other, or that one format is affected by an added construct (McDonald, 2002). One way of determining correlation between rank-ordered data is with Kendall's Tau. Kendall's Tau is a non-parametric correlation that is performed on ranked data, measuring the proportion of agreement and disagreement in pairs of data and scaled from -1 to 1, just like any other correlation (Kendall & Gibbons, 1990). Based on the results of this study, there is a significant difference between the rank-order based on PBT and CBT at fourth grade ( $r_{\tau} = .50, p < .0001$ ) and fifth grade ( $r_{\tau} = .39, p < .0001$ ). At fourth grade, half of the students experienced a rank order change of over 10 places. At fifth grade, one-third of the students experienced a rank order change of over 20 places.

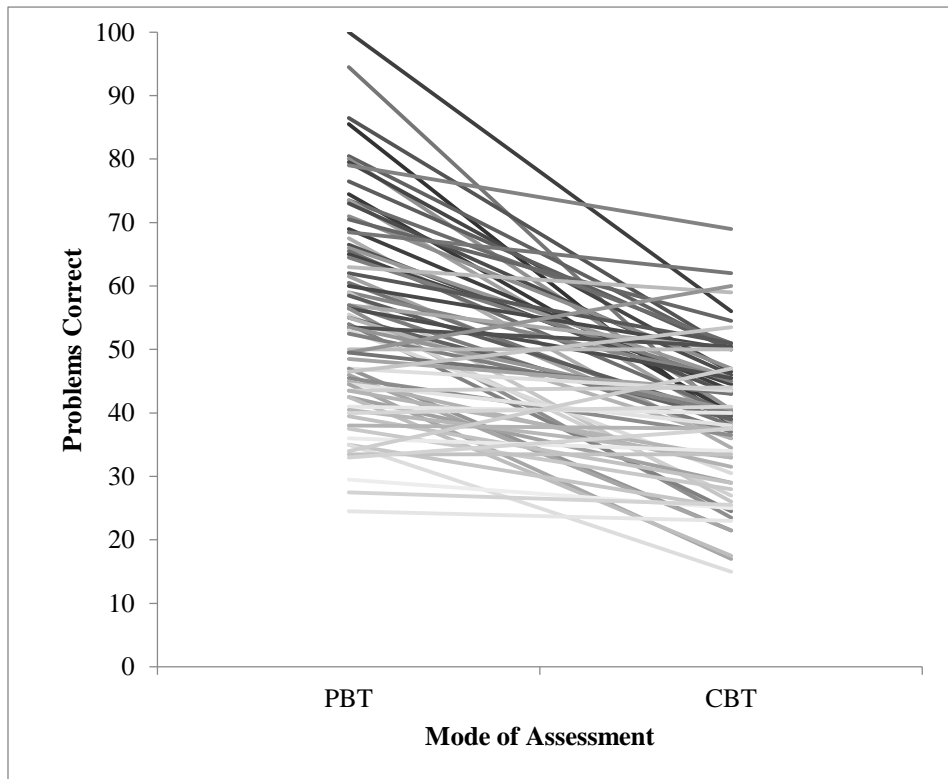
**Figure 8**

*Ranking of Individuals on PBT and CBT Modes- Fourth Grade*



**Figure 9**

*Ranking of Individuals on PBT and CBT Modes- Fifth Grade*





## **Correlate with Criterion Measure**

Additional correlations were computed with performance on different assessment modes and an external assessment of mathematical ability. The purpose of this is to determine if both modes of assessment correlate similarly with an assessment of similar content, also known as criterion-related validity. The GMADE was completed as a criterion measure of mathematical ability. Fourth and fifth grade correlations were calculated separately, as they did not all complete the same version of the GMADE, but completed the GMADE form for the appropriate grade level. Meng's Z-test was used to determine if overlapping correlations with the GMADE were significantly different (Meng, Rubin, & Rosenthal, 1992). At the fourth grade level, both modes correlated similarly with the GMADE assessment. The PBT mode and GMADE assessment had a correlation of  $r = .52$  ( $p < .01$ ), and the CBT mode and GMADE assessment had a correlation of  $r = .54$  ( $p < .01$ ). The difference in correlations is not significant (Meng's  $Z = .27$ ,  $p = .79$ ). At the fifth grade level, the PBT mode had a lower correlation than the CBT counterpart. The PBT mode and GMADE assessment had a correlation of  $r = .26$  ( $p < .05$ ), and the CBT mode and GMADE assessment had a correlation of  $r = .44$  ( $p < .01$ ). The differences in correlation for fifth grade are not quite statistically significant (Meng's  $Z = 1.72$ ,  $p = .08$ ).

## **Comparable Means and Standard Deviations**

As shown in Table 4, the overall mean for the paper-based administration was 48.16 problems correct. For the computer-based administration, the overall mean was 35.25 problems correct. See Table 4 for the breakdown for means at 4<sup>th</sup> and 5<sup>th</sup> grades.

A dependent samples *t*-test was used to analyze the difference between mean scores on paper-based and computer-based tests of fact fluency. The dependent *t*-test assumes that the sampling distribution is normal. With a sample of 145 students, it is assumed that the distribution is normal, as verified by kurtosis and skewness values all less than 1.0. Skewness and kurtosis values greater than 1.0 are considered questionable, and values above 2.0 problematic when assessing normality (Tabachnick & Fidell, 2012). See Table 4 for descriptive statistics.

The purpose of the dependent samples *t*-test was to determine if the difference between PBT and CBT performance was statistically significant. On average, participants performance as determined by problems correct was significantly greater on paper-based tests ( $M = 48.16, SE = 1.45$ ) than on computer-based tests ( $M = 35.25, SE = .89$ )  $t(144) = -12.18, p = .000, r = .71$ .

**Table 4**

*The Distribution of Problems Correct on PBT and CBT M-CBM Probes*

Test Mode	Min	Max	Range	Mean	SD	Skew	Kurtosis	SE
All	7	100	93	48.16	17.42	.35	.02	1.45
PBT 4th	7	78	71	40.13	14.81	.28	.15	1.77
5th	25	100	75	55.85	16.29	.41	-0.21	1.89
All	12	69	57	35.25	10.77	0.29	-0.02	0.89
CBT 4th	12	53	41	30.87	8.41	0.07	-0.01	0.99
5th	15	69	54	39.46	11.14	-0.01	-0.10	1.29

## Findings for Research Question 2

Are there differential mode effects on computer-based tests based on sex, grade level, or ability level?

## Sex

More males than females were represented in this study. Females comprised 45% ( $n = 66$ ) with males making up 55% ( $n = 79$ ) of the total. Using Levene's test for equality of variance, sex met the equality of variance assumption with PBT ( $F = .918, p = .340$ ) and CBT ( $F = .284, p = .595$ ). A mixed design analysis of variance was conducted to compare performance on M-CBM multiplication probes between males and females across the paper-based and computer-based modes of assessment. Table 5 shows the descriptive statistics for the PBT and CBT performance across male and female groups. Results indicated there was no significant interaction between sex and mode of assessment ( $F(2, 142) = .263, p = .609$ ).

**Table 5**

*Descriptive Statistics for Performance of Male and Female Participants Across PBT and CBT Modes of Assessment*

Mode	Male			Female		
	M	SD	N	M	SD	N
PBT	47.18	16.97	79	49.33	18	66
CBT	34.78	11.02	79	35.83	10.50	66

## Grade

A mixed design analysis of variance was conducted to compare performance on M-CBM multiplication probes between fourth and fifth grades across the PBT and CBT modes of assessment. Using Levene's test for equality of variance, grade level met the equality of variance assumption with PBT ( $F = 1.581, p = .211$ ), but indicated unequal variances in CBT ( $F = 4.442, p = .037$ ), so degrees of freedom were adjusted from 142 to 136. Results indicate a significant interaction between grade level and mode of assessment ( $F(2,136) = 12.169, p = .001$ ). When comparing performance between fourth

and fifth grades, the results indicate the CBT format has more of an influence on fifth grade performance than for fourth grade as shown in the differences in means and standard deviations: Refer to Table 4 for descriptive statistics for PBT and CBT performance across grade levels.

### **Overall Mathematics**

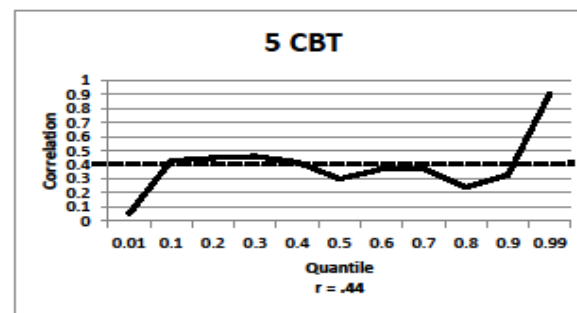
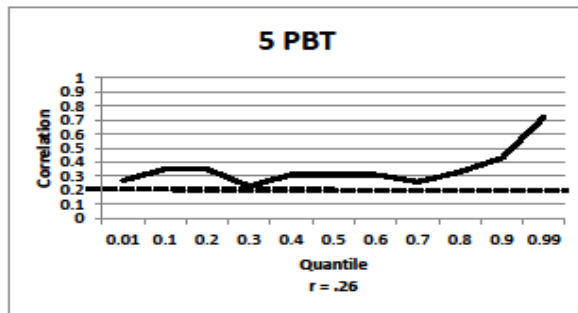
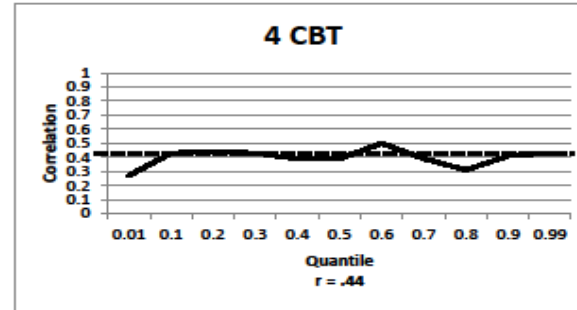
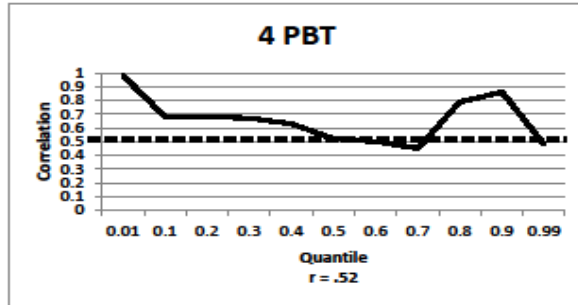
Participants completed the GMADE assessment as a measure of general mathematical ability. Comparisons have already been made between performance on the PBT, CBT, and GMADE in the sections describing the results for correlations with a criterion measure in the previous section.

Quantile regression curves were used further examine the relation between performance on the paper- or computer-based assessment and the GMADE. Quantile regression allows for the examination of two variables across the entire distribution of scores, instead of an estimation based on the mean of the entire group. Figure 5 shows four quantile regression plots. The overall correlation is marked as a dotted line across each plot. There are 11 quantile correlations in a plot, each representing a quantile (.01, .1, .2, .3... .99). The shape of the curve reflects the relation between the format of the assessment (PBT or CBT) and the performance on GMADE. For all four plots, the shape of the curve was relatively flat, indicating a similar strength of relation between the format of the assessment and performance on the GMADE across the distribution. The magnitude of the correlation in all four plots ranged from weak to moderate ( $r = .22$  to  $r = .52$ ). These findings indicate that overall mathematics ability, as determined by the GMADE, does not contribute to differences in performance on PBT and CBT M-CBM probes.

Visual examination does show some differences in correlation at the extreme ends of the distribution, but this can be misleading. There are fewer data points making up the high and low ends of the quantiles, which leads to points that fluctuate more than the center points of the distribution. In particular, the .01 and .99 quantiles are rather spurious, and most likely reflect scores of just a few students. Group sizes ( $n = 69$  and  $n = 73$ ) also contributed to fluctuating estimates. Larger groups sizes would have increased the stability and provided smoother plot lines (Koenker, 2005).

**Figure 10**

*Quantile Plots*



## **Summary of Results**

Two research questions were examined in this study. The first research question examined the comparability of scores on M-CBM probes administered through PBT and CBT modes. Findings were that the mode of assessment had a significant effect on scores, including significant differences in means, score distributions, rankings, and correlations. The second research question examined how the factors of sex, grade, and general mathematical ability contribute to differences in PBT and CBT modes of assessment. Findings were that sex was not a significant factor, but grade was a significant factor. Mathematical ability remained consistent and was not seem a factor in determining performance.

## **CHAPTER FIVE**

### **DISCUSSION**

The purpose of this study was to examine the differences in student performance when given identical paper-based and computer-based M-CBM probes and to investigate the differences that might be accounted for by the factors of sex, age, and overall ability in mathematics. The study was designed to answer the following research questions: (1) Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer? (2) Are there differential mode effects on computer-based tests based on sex, grade level, or ability level?

A mixed-factorial design with both within- and between-subject variables was used to investigate the differences between performance on paper-based and computer-based tests of mathematical fluency. Participants completed both paper- and computer-based tests, as well as the Group Math Assessment and Diagnostic Evaluation as a measure of general math ability.

#### **Research Question 1**

Are there differences in fluency-based performance on math computation problems presented on paper versus on the computer?

Results indicate that the scores from identical tests of fluency administered using paper or computer are not comparable. Following the guidelines set forth by the International Testing Commission (2006) on establishing comparability, it is necessary to show that the PBT and CBT versions

- Have comparable reliabilities.
- Correlate with each other.



- Correlate similarly with an external criterion measure.
- Generate comparable means and standard deviations or can be appropriately adjusted to provide comparable scores.

The data collected in this study were examined following these guidelines. Parallel-forms reliability between the two versions of the PBT and CBT were both over .80, which is an indication that both versions are parallel, however in comparing the reliabilities, they are not comparable. This is the first step and must be in place before examining additional comparability considerations, but it is a step missed by many conducting research in this area. It is crucial that any assessment that will be used to make decisions regarding student success, placement, or instruction adhere to good practices for test development (Coyne & Bartram, 2006). This would be the expectation even if when not examining for comparability.

After determining both sets of forms show an acceptable level of reliability within each format, comparisons were made between the test formats. Overall, PBT and CBT formats show a correlation of 0.68, which does not meet the acceptable level of .80 for reliability (Center on Response to Intervention, n.d.). This was the first indication that the two formats are not comparable. When separated out by grade level, the correlation between formats in fourth grade is stronger at 0.73, compared to the fifth grade correlation between formats (0.54). These grade level correlations give the first indication that grade may have some effect on the comparability of scores.

At this point, the requirement of correlation between formats was not met in this study, but it is informative to look at the correlation between formats in a different way. Figures 8 and 9 show the rank order of individual students. If the two formats were

highly correlated, the expectation would be that rank order would not change much between the two formats. If student ranking changes from one test to the other, this could be an indication that there is some other skill or attribute that is contributing to performance on one or both sets of assessments. In fourth grade, the first, last, and three students in the middle achieved a consistent ranking across modes, but all of the other students changed in the rankings. In fifth grade, there were no students who stayed at the same rank. Appendices E and F give the rankings and scores of individual students at fourth and fifth grade. Statistical analysis (Kendall's Tau) shows that the differences in rank order are statistically significant. Of the 71 fourth graders, five students maintained rank, half of the fourth grade students saw changes of less than 10 places, and 30 student rankings were changed by more than 10 places. The changes at fifth grade were more pronounced. Of the 74 students, none retained rank and over two-thirds were changed by more than 10 places. This is yet another finding that indicates a difference in performance based on grade level. Even students who obtained the same score on the paper- and computer-based tests saw a change in rank. For example, a fifth grade student scored 50 problems correct on paper and ranked 44<sup>th</sup>, but 50 problems correct on the computer ranked 11<sup>th</sup>. This inconsistency in rankings and low correlation between formats suggests that there is some construct irrelevant variance involved, and that some other skill or attribute is contributing to performance on one or both modes of assessments. Further examination of the change in rankings also shows that while most students were faster on paper, some were faster on the computer, and a few displayed consistent performance across modes. This shows that an examination of the role of individual student differences would be helpful and should be included in future studies.

The third factor in considering comparability between formats includes an examination of correlations with a criterion measure. Using the GMADE as an external assessment of mathematics, the correlation of both formats at fourth grade was very similar. The paper-based mode and the GMADE had a correlation of 0.52 and the computer-based mode and the GMADE had a correlation of 0.54. Using Meng's Z transformation, this difference was not significant. For fifth grade students, the correlation was not similar. The PBT mode and GMADE had a correlation of 0.26, while the CBT and GMADE had a correlation of 0.44. This was an almost significant difference in correlation. This difference in correlations aligns with the results on reliability and correlation and suggests that the differences in fifth grade are more pronounced than the differences in fourth grade.

The final piece in determining comparability is to analyze comparabilities in means and score distributions. Performance was significantly higher on PBT than CBT, so comparable means were not found between the two formats. The difference in scores also increased from fourth to fifth grade.

None of the reviewed studies used the International Testing Commission guidelines exactly as they are prescribed. Perhaps this is because when the guidelines were initially published, they were aimed at test publishers and developers, and it was at a time where the move to CBTs was focused on large-scale assessments. Most of the existing research comparing paper- and computer-based tests focus analysis on deciding comparability based on means and distribution only. Less than half of the reviewed comparability studies included data analysis beyond an examination of means. This

indicates a need for more research in this area, as the current body of research has not fully examined the concept of comparability beyond equal means.

It is important to note that even with different means, it would be possible to consider them as equivalent if they could be resolved through statistical methods, such as linear transformation or equipercentile equating. This is not possible for this data set, as the low correlation between formats and inconsistencies in rank order would not allow for any linear transformation to work (McDonald, 2002).

### **Research Question 2**

Are there differential mode effects on computer-based tests based on sex, grade level, or ability level? Results indicate that females outperform males, but it is not statistically significant. Based on grade, there is a significant difference in performance at fourth and fifth grade. When comparing performance on PBT and CBT M-CBMs between fourth and fifth grades, the results indicate the CBT format has more of an influence on fifth grade performance than for fourth grade. This difference in performance based on grade level is strengthened by the findings for the first research question, where correlations between PBT and CBT formats and correlations with the criterion measure (GMADE) differed from fourth grade to fifth grade.

Two of the overall findings of this study reflect the results of previous studies. First, in this study, performance was a measure of speed, which was significantly decreased by the introduction of the CBT format. Multiple studies have found that computerized assessments had a significant effect on speed. Bodmann and Robinson (2004) found that the difference in speed was due to the ability to go back and change answers on one of the CBT versions, as did Goldberg and Pedulla (2002). In the current

study, students could go back and change answers, but were not instructed to do so.

Second, no difference was found based on sex. When examining multiple choice course exams, Clariana and Wallace (2002) found no difference between males and females.

Similar results were found in a study of the NAEP (Horkay et al., 2006) and a state level assessment (Poggio et al., 2005). The last two findings of the current study do not reflect what has been found in previous research. First, the current study shows that the CBT version has a significant effect on performance on fifth grade compared to fourth grade.

There are very few studies examining differences across grade levels, and the only reviewed study actually found the opposite. Pomplun and Custer (2005) examined how the effect of PBT and CBT modes of a reading assessment changed from kindergarten to third grade. They found that there were differences favoring the PBT mode, but the differences between PBT and CBT scores decreased as the grade increased. Pomplun and Custer (2005) felt that this decrease in score differences was due to an interaction of practice effects and familiarity with the content. This could also be a possible explanation for the differences in grade level performance in the current study. Fifth grade students have had an additional year of practice with the PBT format, allowing them to be even more familiar and skilled in both content and format. The introduction of the CBT format highlights the unfamiliarity with the format, leading to a difference in scores. Second, in the current study, there was no difference in performance on PBT and CBT found based on overall performance in mathematics, but Clariana & Wallace (2002) found a mode effect based on the factor of content familiarity.

## **Limitations**

Three limitations have been identified for this study. First, the CBT format was new for all students. Students had to use the Tab key to move from one problem to the next, and enter answers using the keypad or number line. They were given time to practice on the CBT format, but all students were already familiar with the PBT format and have been exposed to that type of test for determining fluency for multiple years. This could result in a threat to internal validity due to testing (Shadish, Cook, & Campbell, 2002). Students have had multiple opportunities over the years to practice using the PBT format. The difference in performance across modes could simply be due to a lack of opportunity on both formats.

Second, no data were collected on level of computer use. As noted earlier, self-reported data on computer use is often unreliable, therefore it was a purposeful decision to not include it in this study. This type of data would be helpful when examining the differences in performance, as it would be one piece of information that might help explain individual differences.

Third, sample size of students at each grade level was a limitation when conducting the quantile regression. This can be seen in the variability at the high and low ends of the distribution, and larger group sizes would have made for a more stable plot. The conclusions drawn about the correlations between different levels of overall mathematical ability and performance on PBT and CBT formats may have been affected by smaller sample sizes. Despite these potential limitations, there are implications from the findings.

## **Implications for Practice**

This study of paper- and computer-based modes of assessment is important because it examines mode effects using M-CBMs that can be used to make instructional decisions. In many cases, teachers have options in choosing paper-based tests or computer-based tests in their classrooms, and technologically shrewd teachers can even create their own computer-based assessments for classroom use. Because of this, continuing research in this area, particularly at the classroom level, is necessary. From a classroom teacher's perspective, the use of CBTs is alluring, as it can cut down on time needed for administration, scoring, and analyzing when compared to PBTs.

In terms of the content of this study, one might think that it wouldn't make much of a difference if one classroom uses a paper-based test of mathematical fluency and another classroom uses the same form, but computer-based. In classrooms and schools, these data can be used to determine who needs additional support. As the results from this study have shown, performance levels can vary drastically depending on the format of the M-CBM. Based on CBT performance, the decision might be made to provide additional instruction or practice with basic math skills, when the results obtained from the PBT version would show that the student does not need any extra support in this area. This study shows that those scores are not comparable, and decisions made regarding instruction may be inaccurate, leading to a waste of instructional time and resources.

There may also be implications for practice based on the content and grade level. The skills involved in using a CBT for mathematics may be different from the skills needed to complete a CBT in reading or writing. It is also important to examine how

skills used to complete CBTs change throughout grade levels. These could both be contributing factors in the inconsistencies in existing research.

### **Future Research**

The findings of this study add to the literature base regarding mode effects of assessment. As computer use in the classroom continues to increase, it is necessary to study the use of computer-based classroom assessments, which is lacking in the existing research. Notably absent from the research comparing PBT and CBT formats is research with CBM measures in reading, writing, and mathematics. Considering the characteristics of CBM measures, it is tempting for teachers to develop their own CBTs, or use some of the existing CBT options for CBM. Standardized administration and timing easily translate to a CBT format, as does automatic scoring, feedback, and analysis. Due to the attractiveness of this option, it is particularly important to extend the research on mode effects and CBM in all areas and examine how a change in formats can affect scores on CBMs that teachers use to make daily instructional decisions.

Additional studies are also needed regarding practice with CBT formats. Most of the existing studies compare assessments given on paper and on the computer with limited attempts. Based on individual students and the use of technology in their classrooms, this may have been one of the few opportunities to use a computer for an assessment. A study that allows practice in the CBT format over time would give some idea about how practice may or may not change the difference in scores based on format. Along the same lines, studies with an intervention component focusing on increasing computer and keyboarding skills would be beneficial. This could address the common



question of how the level of computer skill relates to performance on CBT measures, but do so in a way that is more reliable than the surveys used in most of the existing research.

Last, more research is also needed with a wider range of ages. A consistent result from this study was the issue of differing performance based on grade level. It was clear that at least for this particular group of students, the negative effect of the CBT was greater for fifth grade students over the fourth grade students. It is necessary to see how this trend extends to a broader range of grade levels. The findings from other studies indicate that performance differs based on grade level, but the findings are inconsistent, and expanding grade levels would allow for more in depth examination of the trend.

### **Conclusion**

Trends in the use of technology continue to indicate that the use of computer-based tests is not going away. For teachers, CBTs have the potential to save time in administration, scoring, and analysis, which would hopefully lead to increased and consistent use as an instructional decision making tool. For students, CBTs can be motivating. Even when it is not the best option in terms of performance, many students will still choose to use the technology-based format just because they like it (Penuel, 2006). For these reasons alone, technology will stay at the forefront of classroom-based assessment. Instead of fighting against it, it is important that research is conducted on which factors lead to the differences in performance. Once it is determined what the key factors are, research needs to continue on how to compensate for those factors, hopefully allowing for comparable PBT and CBT formats.

Studies in this area have been conducted for decades. It is important that this trend continues. Even though the studies have examined many of the same factors repeatedly, it

is necessary to do so, because of the changes in technology. In the 1990s, studies using large desktop computers were considered current, and issues included screen refresh rate and screen resolution. In the 2000s, the issues with screens were generally resolved, but research still continued in labs filled with slightly smaller desktop computers. At that time, the content was limited to CBTs provided by testing companies, as creating CBTs would have required extensive programming skill. Now that it is much easier to create content, teachers are able to program their own CBTs, and those CBTs can be given on a very fast desktop computer, classroom sets of laptops, tablets, and even phones. The trend of innovation in technology will continue, and therefore this line of study will continue and need to be replicated whenever any new piece of technology is introduced.

## APPENDIX A

### DIRECTIONS FOR PAPER-BASED PROBES

Distribute folders to students. In each folder there are two stapled sets of two sheets each, containing multiplication facts 0-12. The order of the sets in the folder has been randomized, so not all students need to have the same set out at the same time. Have a copy of the test in hand and give the following highlighted instructions out loud to the students:

**Please take out the first stapled set of worksheets, put your name at the top and then turn it over face down on your desk.** Make sure students have just one set of worksheets out. **You are going to do a math worksheet.** (hold up a copy of the test) **This worksheet is made up of many problems. When I say “begin” turn your sheet over and begin solving the problems. Start with the first problem and work across each row from left to right.** (point to the first problem and follow with your finger across the entire row) **If you come to a problem you don’t know, just skip it and go to the next problem. You can go back and try to solve the problems you skipped if you have time. Do as many problems as you can and continue working until I say “stop”. You will have 2 minutes to complete the worksheet. At the end of 2 minutes I will say “Stop”. You will then put your pencils down and turn your sheet over. Are there any questions?** (Answer any questions to make sure everyone understands how to take the test). Then say, **“Begin”** (start the timer).

After 2 minutes- **Stop. Please put the completed worksheet back in the folder and take out the next set. Put your name at the top and then turn it over face down on your desk. You will now do a similar set of problems. When I say “begin” turn your sheet over and begin solving the problems. You will have 2 minutes to complete the worksheet. At the end of 2 minutes I will say “Stop”. You will then put your pencils down and turn your sheet over. Are there any questions?** (Answer any questions to make sure everyone understands how to take the test). Then say, **“Begin”** (start the timer).

Collect the folders.

APPENDIX B

PAPER-BASED PROBE 1 (PAGE 1 OF 3)

Name: \_\_\_\_\_

Student ID Number: \_\_\_\_\_

$$\begin{array}{r} 11 \\ \times 4 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 4 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 10 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 12 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 10 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 5 \\ \times 6 \\ \hline \end{array}$$

$$\begin{array}{r} 0 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 11 \\ \times 9 \\ \hline \end{array}$$

$$\begin{array}{r} 4 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 9 \\ \hline \end{array}$$

$$\begin{array}{r} 2 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 11 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 8 \\ \hline \end{array}$$

$$\begin{array}{r} 3 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 12 \\ \times 12 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 10 \\ \hline \end{array}$$

$$\begin{array}{r} 0 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 7 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 12 \\ \times 7 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 10 \\ \hline \end{array}$$

$$\begin{array}{r} 4 \\ \times 0 \\ \hline \end{array}$$

$$\begin{array}{r} 12 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 10 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 1 \\ \times 8 \\ \hline \end{array}$$

$$\begin{array}{r} 3 \\ \times 9 \\ \hline \end{array}$$

$$\begin{array}{r} 4 \\ \times 6 \\ \hline \end{array}$$

$$\begin{array}{r} 11 \\ \times 4 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 6 \\ \hline \end{array}$$

## APPENDIX C

### DIRECTIONS FOR COMPUTER-BASED PROBES

**You will be completing math problems on the computer. For Class ID, please enter your class ID and click Submit. For Student ID, enter your ID number.**

**We will try a sample test first. Click next to Sample, then click submit. Click in the first box. Use the number keys to answer the question, then use the Tab key to go to the next problem. If you come to a problem you don't know, hit the tab key to go to the next problem. Go ahead and try the problems on this page. *When the students are done, prompt them to click the submit button.***

**Now we will do 2 timed multiplication tests. Where it says Select Test, click A. You will have 2 minutes to complete the first set of problems. When I say begin, you will start with the first problem and work across each row, using the number key to enter your answer and the tab key to move to the next problem. Do as many problems as you can and continue working until I say stop. You will have 2 minutes to complete as many problems as you can. Are there any questions? Begin *(start the timer- 2 minutes)*.**

**Stop. Scroll down to the bottom of the page and click submit.**

**Where it says Select Test, click B. You will have 2 minutes to complete as many problems as you can. Are there any questions? Begin.**

**Stop. Scroll down to the bottom of the page and click submit.**

APPENDIX D

COMPUTER-BASED PROBE 1 (PAGE 1 OF 3)

$$\begin{array}{r} 11 \\ \times 4 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 4 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 10 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 12 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 10 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 5 \\ \times 6 \\ \hline \end{array}$$

$$\begin{array}{r} 0 \\ \times 11 \\ \hline \end{array}$$

$$\begin{array}{r} 11 \\ \times 9 \\ \hline \end{array}$$

$$\begin{array}{r} 4 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 9 \\ \hline \end{array}$$

$$\begin{array}{r} 2 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 11 \\ \times 1 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 8 \\ \hline \end{array}$$

$$\begin{array}{r} 3 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 12 \\ \times 12 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 10 \\ \hline \end{array}$$

$$\begin{array}{r} 0 \\ \times 2 \\ \hline \end{array}$$

$$\begin{array}{r} 9 \\ \times 7 \\ \hline \end{array}$$

$$\begin{array}{r} 7 \\ \times 3 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 5 \\ \hline \end{array}$$

$$\begin{array}{r} 6 \\ \times 11 \\ \hline \end{array}$$

## APPENDIX E

### FOURTH GRADE PARTICIPANT RANKING

*Fourth Grade Individual Ranking by PBT  
and CBT*

PBT	PBT Rank	CBT	CBT Rank
77.5	1	53	1
72	2	42	7
71.5	3	49.5	2
68.5	4	33.5	24
66.5	5	37.5	16
62.5	6	39	14
62	7	27	49
60.5	8	40.5	10
59.5	9	32	31
59	10	42.5	6
58	11	39.5	12
57.5	12	32	31
51.5	13	41	9
51	14	20	65
48.5	15	38	12
48.5	15	39.5	15
47.5	17	31	38
47	18	31.5	33
47	18	31.5	33
45.5	20	29	27
45.5	20	33	44

45	22	28	46
44.5	23	43.5	4
44	24	31.5	33
43.5	25	24	54
43	26	31.5	22
43	26	34	33
42.5	28	40.5	10
41.5	29	34.5	20
41.5	29	25	27
41.5	29	33	53
41	32	34	18
41	32	36	22
41	32	28.5	45
39.5	35	34.5	20
39.5	35	30.5	39
39.5	35	29.5	42
39	38	23	57
38.5	39	32.5	29
38.5	39	30.5	39
38	41	29.5	42
37.5	42	42	7
37	43	33.5	24
36.5	44	36	18
36	45	45	3
35	46	30	41
34.5	47	33.5	24
34	48	37	5



34	48	43	17
33.5	50	31.5	29
33.5	50	32.5	37
32.5	52	26	51
32.5	52	22	63
32	54	26.5	50
31.5	55	27.5	47
29.5	56	27.5	47
29.5	56	23	57
28.5	58	22.5	61
26.5	59	24	54
24	60	23	57
24	60	21	64
23.5	62	24	54
21.5	63	22.5	61
21.5	63	17.5	67
21	65	23	57
20	66	25.5	52
19	67	20	65
17	68	15	68
16	69	14	70
8.5	70	14.5	69
7	71	12	71

---

## APPENDIX F

### FIFTH GRADE PARTICIPANT RANKING

*Fifth Grade Individual Ranking by PBT and CBT*

<b>PBT</b>	<b>PBT Rank</b>	<b>CBT</b>	<b>CBT Rank</b>
100	1	56	5
94.5	2	39	40
86.5	3	50	11
85.5	4	38	43
80.5	5	50	11
80	6	40.5	33
79.5	7	46.5	19
79	8	69	1
76.5	9	50	11
74.5	10	40	35
73.5	11	51	8
73.5	11	38.5	42
73	13	46	20
71	14	47	17
70.5	15	54.5	6
69	16	43.5	27
68.5	17	62	2
67.5	18	34.5	52
66.5	19	43	30
66	20	45.5	21
65.5	21	40	35
65	22	44.5	25

64.5	23	45	24
63	24	59	4
62	25	51	8
61.5	26	36	50
60.5	27	38	43
60	28	50	11
59	29	45.5	21
59	29	26	63
58.5	31	39.5	39
57	32	50	11
57	32	37	48
56.5	34	45.5	21
56.5	34	23.5	68
55.5	36	27	62
55	37	39	40
55	37	36	50
54	39	24.5	67
53.5	40	50.5	10
53.5	40	40.5	33
53	42	30.5	58
52.5	43	40	35
50	44	50	11
49.5	45	60	3
49.5	45	43	30
48.5	47	43.5	27
47	48	43.5	27
47	48	21.5	70

46.5	50	53.5	7
46	51	21.5	70
45.5	52	36.5	49
45	53	31.5	57
44.5	54	38	43
44.5	54	17	73
43.5	56	44	26
43.5	56	33	55
42.5	58	29	59
42.5	58	17.5	72
41	60	40	35
40.5	61	33	55
40	62	41	32
39.5	63	29	59
38	64	37.5	46
37.5	65	28	61
36	66	34	53
35	67	25	65
35	67	15	74
34	69	47	17
33.5	70	33.5	54
33	71	37.5	46
29.5	72	25	65
27.5	73	25.5	64
24.5	74	23	69

---

## REFERENCES

- 1:1 schools in Iowa. (2014). Retrieved from <https://www.aea267.k12.ia.us/techintegration/current-initiatives/one-to-one/1-to-1-schools-in-iowa/>
- American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baldwin, F. (1999). Taking the classroom home. *Appalachia*, 32(1), 10-15.
- Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, 28(2), 117-148.
- Bebell, D., & Kay, R. (2010). One to one computing: A summary of the quantitative results from the Berkshire wireless learning initiative. *The Journal of Technology, Learning and Assessment*, 9(2).
- Bebell, D., O'Dwyer, L. M., Russell, M., & Hoffmann, T. (2010). Concerns, considerations, and new ideas for data collection and research in educational technology studies. *Journal of Research on Technology in Education*, 43(1), 29-52.
- Becker, J. D. (2006). Digital equity in education: A multilevel examination of differences in and relationships between computer access, computer use and state-level technology policies. *Education Policy Analysis Archives*, 15(3), 1-38.
- Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In M. vonDavier et al. (Eds.), *The Role of International Large-Scale Assessments: Perspective from Technology, Economy, and Educational Research* (pp. 25-45). Springer Science.
- Bennett, R.E. (2003). *Online assessment and the comparability of score meaning*. Princeton, NJ: Educational Testing Service.
- Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, 39(1), 370-407.
- 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. *The Journal of Technology, Learning and Assessment*, 6(9).

- Bernard, M., Liao, C. H., & Mills, M. (2001, March). The effects of font type and size on the legibility and reading time of online text by older adults. In *CHI'01 extended abstracts on Human factors in computing systems* (pp. 175-176).
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, 86(1), 8-21.
- 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. (2009). Experiences from large-scale computer-based testing in the USA. In F. Scheurmann & J. Bjornsson (Eds.), *The Transition to Computer-Based Assessment* (pp. 39-44). Luxembourg, Office for Official Publications of the European Communities.
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5-12.
- Brown, J., Hinze, S., & Pellegrino, J. W. (2008). Technology and formative assessment. In T. Good (Ed.), *21<sup>st</sup> Century Education. Vol 2. Technology* (pp. 245-255). Thousand Oaks, CA:Sage.
- Bryant, D. P., Hartman, P., & Kim, S. A. (2003). Using explicit and strategic instruction to teach division skills to students with learning disabilities. *Exceptionality*, 11(3), 151-164.
- Bugbee Jr, A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, 23(1), 87-100.
- Choi, S. W. & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. Paper presented at the 2002 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Christ, T. J. & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, 33(3), 387-400.
- Christ, T. J., Scullin, S., Tolbize, A. & Jiban, C. L. (2008). Curriculum-based measurement of math computation. *Assessment for Effective Intervention*, 33(4), 198-205.
- Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice*, 42, 75-83.

- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33(5), 593-602.
- Clark, L. S., Demont-Heinrich, C., & Webber, S. (2005). Parents, ICTs, and children's prospects for success: Interviews along the digital "Access Rainbow". *Critical Studies in Media Communication*, 22(5), 409-426.
- Clarke, B. & Shinn, M.R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33(2), 234-248.
- Common Core State Standards Initiative. (2010). *Common core state standards for mathematics: Grades K-8*. Retrieved from <http://www.corestandards.org/Math/>.
- Coyne, I., & Bartram, D. (2006). Design and development of the ITC guidelines on computer-based and Internet-delivered testing. *International Journal of Testing*, 6(2), 133-142.
- Dean, V. & Martineau, J. (2012). A state perspective on enhancing assessment and accountability systems through systematic implementation of technology. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments* (pp. 25-53). Charlotte, NC; Information Age Publishing, Inc.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37(3), 184-192.
- Duhon, G.J., House, S.H., & Stinnett, T.A. (2012). Evaluating the generalization of math fact fluency gains across paper and computer performance modalities. *Journal of School Psychology*, 50(3), 335-345.
- Eaves, R. C., & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. *The Journal of Experimental Education*, 55(1), 23-26.
- Fletcher, G. H. (2010). Race to the top: No district left behind: The reforms elicited by the competition that will likely affect your school system, whether you like it or not. *THE Journal (Technological Horizons in Education)*, 37(10), 17.
- Foegen, A. & Deno, S.L. (2001). Identifying growth indicators for low-achieving students in middle school mathematics. *The Journal of Special Education*, 35(1), 4-16.

- Foegen, A. (2005). *Developing measures in algebra*. Presentation at the annual Pacific Coast Research Conference: San Diego.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education, 41*, 121–139.
- Foegen, A., Olson, J.R., & Impeccoven-Lind, L. (2008). Developing progress monitoring measures for secondary mathematics: An illustration in algebra. *Assessment for Effective Intervention 33*(4), 240-249.
- Fuchs, L. S. (1992). Classwide decisionmaking with computerized curriculum-based measurement. *Preventing School Failure: Alternative Education for Children and Youth, 36*(2), 30-33.
- Fuchs, L. S. (1998). Computer applications to address implementation difficulties associated with curriculum-based measurement. In Shinn, M. R. (Ed.), *Advanced applications of curriculum-based measurement*. (pp 89-112). New York: Guilford Press.
- Fuchs, L. S. (2004) : The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.
- Fuchs, L. S., & Fuchs, D. (2001). Principles for the prevention and intervention of mathematics difficulties. *Learning Disabilities Research & Practice, 16*(2), 85-95.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M. (2006). The cognitive correlates of third grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Education Psychology, 98*, 29-43.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1994). Strengthening the connection between assessment and instructional planning with expert systems. *Exceptional Children, 61*, 138-146.
- Fuchs, L.S., Fuchs, D., Hamlett, C. L., & Allinder, R.M. (1989). The reliability and validity of skills analysis within curriculum-based measurement. *Diagnostique, 14*, 203-221.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Hasselbring, T. (1987). Using computers with curriculum-based progress monitoring: Effects on teachers effectiveness and satisfaction. *Journal of Special Education Technology, 8*(4), 14-27.



- Fuchs, L.S., Fuchs, D., Hamlett, C.L., & Stecker, P.M. (1991). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review*, 19(1), 6-23.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education*, 33(1), 18-28.
- Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38(4), 293-304.
- Goo, M., Watt, S., Park, Y., & Hosp, J. (2012). A guide to choosing web-based curriculum-based measurements for the classroom. *Teaching Exceptional Children*, 45, 34-40.
- Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement*, 62(6), 1053-1067.
- Gould, J. D., Alfaro, L., Finn, R., Haupt, B., & Minuto, A. (1987). Reading from CRT displays can be as fast as reading from paper. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(5), 497-517.
- Gray, L., Thomas, N., & Lewis, L. (2010). *Teachers' Use of Educational Technology in U. S., Public Schools: 2009* (NCES 2010-040). National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Washington, DC.
- Halverson, R., & Shapiro, R. B. (2012). Technologies for education and technologies for learners: How information technologies are (and should be) changing schools. *Wisconsin Center for Educational Research (WCER), Working Paper*, 6.
- Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: does the medium in which assessment questions are presented affect children's performance in mathematics? *Educational Research*, 46(1), 29-42.
- Harrison, C., Lunzer, E. A., Tymms, P., Fitz-Gibbon, C. T., & Restorick, J. (2004). Use of ICT and its relationship with performance in examinations: a comparison of the ImpaCT2 project's research findings using pupil-level, school-level and multilevel modelling data. *Journal of Computer Assisted Learning*, 20(5), 319-337.

- Hippisley, J., Douglas, G., & Houghton, S. (2005). A cross-cultural comparison of numeracy skills using a written and an interactive arithmetic test. *Educational Research*, 47(2), 205-215.
- Holland, J. & Holland J. (2014). Implications of shifting technology in education. *TechTrends*, 58(3), 16-25.
- Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does It Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment*, 5(2), n2.
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABC's of CBM: A practical guide to curriculum-based measurement*. New York: Guilford.
- International Test Commission. (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing*, 6(2), 143-171.
- Jiban, C. L., & Deno, S. L. (2007). Using Math and Reading Curriculum-Based Measurements to Predict State Mathematics Test Performance Are Simple One-Minute Measures Technically Adequate? *Assessment for Effective Intervention*, 32(2), 78-89.
- Kapoor, S., & Welch, C. (2011, April). *Comparability of paper and computer administrations in terms of proficiency interpretations*. Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. Retrieved from <https://itp.education.uiowa.edu/ia/documents/Comparability%20of%20Paper%20and%20Computer%20Administrations%20in%20Terms%20of%20Proficiency%20Interpretations.pdf>.
- Kendall, M., & Gibbons, J. D. (1990). *Rank Correlation Methods*. (5<sup>th</sup> ed.). New York: Oxford University Press.
- Kennedy, M. J., & Deschler, D. D. (2010). Literacy instruction, technology, and students with learning disabilities: Research we have, research we need. *Learning Disability Quarterly*, 33(4), 289-298.
- Klein, S. P. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. *Probability and Statistics*, 2, 76-89.
- Koenker, R. (2005). *Quantile Regression*, New York: Cambridge University Press.
- Kolen, M.K. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 73-96.

- Kolen, M. J., & Brennan, R. L. (1995). *Test equating, scaling, and linking: Methods and practices* (2<sup>nd</sup> ed.) New York: Springer.
- Lee, Y. H., Waxman, H., Wu, J. Y., Michko, G., & Lin, G. (2013). Revisit the Effect of Teaching and Learning with Technology. *Educational Technology & Society*, 16(1), 133-146.
- Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, 6(1), 1-24.
- Lembke, E.S., & Foegen, A. (2005). *Creating measures of early numeracy*. Presentation at the Annual Pacific Coast Research Conference. San Diego.
- Lembke, E.S., Hampton, D., & Beyers, S.J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools*, 49(3), 257-272.
- Lembke, E.S. & Stecker, P. (2007). *Curriculum-based measurement in mathematics: An evidence-based formative assessment procedure*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Lim, C. P., Zhao, Y., Tondeur, J., Chai, C. S., & Tsai, C. C. (2013). Bridging the gap: Technology trends and use of technology in schools. *Educational Technology & Society*, 16(2), 59-68.
- Lottridge, S. M., Nicewander, W. A., & Mitzel, H. C. (2011) A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research*, 46, 544-566.
- Luecht, R. M., Hadadi, A., Swanson, D. B., & Case, S. M. (1998). Testing the test: A comparative study of a comprehensive basic sciences test using paper-and-pencil and computerized formats. *Academic Medicine*, 73(10), 51-53.
- McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education*, 39, 299-312.
- McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education*, 41, 68-84.
- Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research*, 24(1), 29-40.
- Meelissen, M. R., & Drent, M. (2008). Gender differences in computer attitudes: Does the school matter? *Computers in Human Behavior*, 24(3), 969-985.

- Meng, X. L., Rosenthal, R., & Rubin, D. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172-175.
- National Center for Education Statistics (NCES). (2011). *NAEP writing computer-based assessment: An overview for Grade 4*. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/about/schools/2012grade4wcbabrochure.pdf>
- National Center on Response to Intervention. (n.d.). Retrieved from <http://www.rti4success.org/>.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. US Department of Education.
- Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352-1375.
- Congress, U. S. (1995). Office of technology assessment. *Protecting privacy in computerized medical information*, 15.
- Organization for Economic Co-Operation and Development. (2007). *PISA 2006 science competencies for tomorrow's world*. Paris, France: Author.
- Parshall, C. G., & Kromrey, J. D. (1993). Computer testing versus paper-and-pencil Testing: An analysis of examinee characteristics associated with mode effect. *Paper presented at the Annual Meeting of the American Educational Research Association*. Atlanta, GA.
- Partnership for Assessment of Readiness for College and Careers. (2013). *Technology guidelines for PARCC assessments*. Washington, DC: Author. Retrieved from <http://www.parcconline.org/sites/parcc/files/Technology%20Guidelines%20for%20PARCC%20Assessments%20v%204%20January%202015.pdf>
- Peak, P. (2005). *Recent trends in comparability studies*. Pearson Educational Measurement.
- Pearson, Inc. (2009). *AIMSweb mathematics concepts and applications administration and technical manual*. San Antonio, TX: Author.
- Pellegrino, J. W. & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*(2). 119-134.
- Penuel, W. R. (2006). Implementation and effects of one-to-one computing initiatives: A research synthesis. *Journal of Research on Technology in Education*, *38*(3), 329-348.

- Phillips, M., Shinn, M. R., & Ditkowsky, B. (2014). The Use of Technology with Curriculum-Based Measurement (CBM). *Academic Assessment and Intervention*, 139.
- 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. *The Journal of Technology, Learning and Assessment*, 3(6).
- Poggio, J. & McJunkin, L. (2012). History, current practice, perspectives and what the future holds for computer based assessment in K-12 education. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments* (pp. 25-53). Charlotte, NC; Information Age Publishing, Inc.
- Pomerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, learning, and Assessment*, 2(6).
- Pomplun, M. & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research*, 32(2), 153-166.
- Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62(2), 337-354.
- Project Tomorrow. (2014). *The new digital learning playbook: Understanding the spectrum of students' activities and aspirations*. Retrieved from <http://www.tomorrow.org/speakup/pdfs/SU13StudentsReport.pdf>
- Quellmalz, E. S. & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323, 75-78.
- Redecker, C., & Johannessen, O. (2013). Changing assessment- Toward a new assessment paradigm using ICT. *European Journal of Education*, 47(1), 79-95.
- Reiser, R. A., & Dempsey, J. V. (2007). *Trends and issues in instructional design and technology*. Upper Saddle River, NJ: Person Merrill Prentice Hall.
- Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, 47, 427-469.
- Ripley, M. (2009). Transformational computer-based testing. In F. Scheurmann & J. Bjornsson (Eds.), *The Transition to Computer-Based Assessment* (pp. 89-91). Luxembourg, Office for Official Publications of the European Communities.

- Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7
- Russell, M., Bebell, D., & Higgins, J. (2004). Laptop learning: A comparison of teaching and learning in upper elementary classrooms equipped with shared carts of laptops and permanent 1: 1 laptops. *Journal of Educational Computing Research*, 30(4), 313-330.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.
- Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention*. (3<sup>rd</sup> ed.). New York, NY: Guilford Press.
- Shapiro, E.S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, 30(2), 15-32.
- Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Sacramento, CA: Author. Retrieved from [http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced\\_Guidelines.pdf](http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced_Guidelines.pdf)
- Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational measurement*, 26(3), 261-271.
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, 42(8), 795-819.
- Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Preventing School Failure: Alternative Education for Children and Youth*, 52(2), 48-58.
- Tabachnick, B., & Fidell, L. (2012). *Using multivariate statistics* (6<sup>th</sup> ed.). Boston, MA: Pearson.
- Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning a second-order meta-analysis and validation study. *Review of Educational research*, 81(1), 4-28.

- Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Education Research Journal*, 31(2), 231-262.
- Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, 31(4), 498-513.
- Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). Computer-Based Testing: Practices and Considerations. Synthesis Report 78. *National Center on Educational Outcomes, University of Minnesota*.
- Vekiri, I., & Chronaki, A. (2008). Gender issues in technology use: Perceived social support, computer self-efficacy and value beliefs, and computer use beyond school. *Computers & Education*, 51(3), 1392-1404.
- Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*, 60(3), 371-384.
- Vispoel, W. P. (2000). Computerized versus paper-and-pencil assessment of self-concept: Score comparability and respondent preferences. *Measurement and Evaluation in Counseling and Development*.
- Vukovic, R. K., & Siegel, L. S. (2010). Academic and cognitive characteristics of persistent mathematics difficulty from first to fourth grade. *Learning Disabilities Research & Practice*, 25(1), 25-38.
- Vryzas, K., & Tsitouridou, M. (2002). The home computer in children's everyday life: the case of Greece. *Journal of Educational Media*, 27(1-2), 9-17.
- Wang, S., Jiao, H., Young, M.J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67(2), 219-238.
- Warschauer, M., Knoble, M., & Stone, L. (2004). Technology and equity in schooling: Deconstructing the digital divide. *Educational Policy*, 18(4), 562-588.
- Waters, J. K. (2012). Resolving the Formative Assessment Catch-22: Teachers Often Have a Hard Time Embedding Assessment in Their Instruction, but Some Technologies Are Making It Easier. *THE Journal (Technological Horizons In Education)*, 39(7), 8.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichdt, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41, 85-120.

- Wheadon, C. (2007). *The comparability of onscreen and paper and pencil tests: no further research required?* Manchester, UK: Centre for Education Research & Policy, Assessment and Qualifications Alliance.
- Wilson, R., Majsterek, D., & Simmons, D. (1996). The effects of computer-assisted versus teacher-directed instruction on the multiplication performance of elementary students with learning disabilities. *Journal of Learning Disabilities*, 29(4), 382-390.
- Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational measurement: Issues and practice*, 8(3), 5-10.
- Woodward, J. (2006). Making reformed based mathematics work for academically low achieving middle school students. *Teaching mathematics to middle school students with learning difficulties*, 29-50.
- Ysseldyke, J. E., & McLeod, S. (2007). Using technology tools to monitor response to intervention. In *Handbook of response to intervention* (pp. 396-407). Springer US.