EXAMINING THE TECHNICAL ADEQUACY OF CURRICULUM-BASED MEASUREMENT PROGRESS MONITORING IN EARLY NUMERACY USING HANDHELD TECHNOLOGY

A Dissertation presented to

the faculty of the Graduate School

University of Missouri

In partial Fulfillment

of the Requirements for the Degree

of Doctor of Philosophy

By

DAVID DEAN HAMPTON

Dr. Erica S. Lembke, Dissertation Supervisor

JULY 2011

The undersigned¸ appointed by the Dean of the Graduate School, have examined the dissertation entitled:

EXAMINING THE TECHNICAL ADEQUACY OF CURRICULUM-BASED MEASUREMENT PROGRESS MONITORING IN EARLY NUMERACY USING HANDHELD TECHNOLOGY

Presented by David Dean Hampton, a candidate for the degree of Doctor of Philosophy and hereby certify that in their opinion it is worthy of acceptance.

____

Dr. Erica S. Lembke

____

Dr. Janine Stichter

____

Dr. Melissa Stormont

____

Dr. Delinda vanGarderen

____

Dr. Cheryl Offutt

DEDICATION

This dissertation is dedicated to my beautiful niece¸ Rebecca Embree, who has served as an inspiration to me every day. Becky, you have shown me that everything is possible as long as you remain positive. I also include in this dedication my mother and father who always allowed me to take risks and to make mistakes, yet they were always there to help me regroup and move forward, always moving forward!

ACKNOWLEDGEMENTS

Sitting and contemplating how to acknowledge all of the important people who have contributed to my doctoral program, is quite a humbling experience. I must begin with Dr. Erica Lembke for sharing your time and wealth of knowledge with me for over six years. I have no comprehension of the right words to express my gratitude to you for your patience, kindness, generosity, and wonderful sense of humor. You have been the best mentor I could have ever imagined. I must also include my sincerest gratitude to the esteemed members of my committee. Dr. Janine Stichter, you have been a guiding force on my journey from the beginning. I still consider helping you with your textbook in 2005 to be one of the best experiences of my time at MU. Dr. Melissa Stormont, you have guided me with your gentle spirit and incredible knowledge, and I will never be able to adequately express my gratitude to you. Dr. Cheryl Offutt, you have been with me during my journey from the beginning also, and I have enjoyed our talks and your experience and example will make me a much better collaborator in the future. Dr. Delinda vanGarderen, I have learned so much from you over the past four years. You have a dedication to your work that is truly inspirational, and I have learned to temper my zeal and consider issues with a more open-mind. I hope that my future work in mathematics instruction and research will meet the standards that you have set.

There are some other important people that I must acknowledge as well. Dr. Timothy Lewis has been the most kind and thoughtful mentor that anyone could have. I feel truly blessed that I was able to be your student and that you never closed your door to my endless interruptions and questions. Janice Jacobson, you have been a true godsend in that you were never too busy to answer a question or to just sit and listen to the droning of an overwhelmed graduate student.

You are the heart of this department and you have made my administrative experiences at MU such a pleasure.

I must acknowledge those who made this dissertation study possible. First, to the administration, faculty and staff at Southern Boone County Primary School; thank you for being such wonderful partners and colleagues. Second, the students who participated in our study; thank you so much for assisting us in our venture and for your enthusiasm, humor, and efforts, you are all superstars. I need to send special thanks to our data collectors, who worked tirelessly, with passion and skill, and a lot of humor. Going through this process with each of you has been a highlight of my professional life.

To my family, I must say thank you for never questioning my desire to pursue higher education. I know that my parents and my brother Patrick are watching me from up above and I pray that I have made them proud. To Cheryl, Bob, and Sheila, thank you for helping me time and time again, and for allowing me to obsess and whine about anything ad nauseam. To my fellow grad students and friends, especially Sushama, Stacey, and Nick; thank you for walking the journey with me, and I know you each will be the most inspiring researchers around.

I have learned enough not to think of this as the culmination of my journey. I am merely standing at the starting line and I want to be like each of you when I finish!

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

EXAMINING THE TECHNICAL ADEQUACY OF CURRICULUM-BASED MEASUREMENT PROGRESS MONITORING IN EARLY NUMERACY USING HANDHELD TECHNOLOGY

David Dean Hampton

Dr. Erica S. Lembke, Dissertation Supervisor

ABSTRACT

The purpose of this study was to examine the use of six early numeracy measures to monitor the mathematics progress of kindergarten and first grade students across 13 weeks. Seventy one kindergarteners were administered oral counting, number identification, missing number, and quantity discrimination measures weekly for 13 weeks. Simultaneously, seventy five first grade students were administered oral-counting, number identification, missing number, next number, number facts, and quantity discrimination measures weekly for 13 weeks. All data was collected via PDA Palm Pilot handheld technology with web-based data management supplied by Wireless Generation, Inc. (mclass:Math software). Alternate form reliability was adequate for instructional decision making on some measures, and low reliability was reported for quantity discrimination, as well as for the next number and number facts measures. Concurrent criterion validity coefficients comparing the measures with student performance on a standardized assessment resulted in weaker coefficients as compared to previous studies that have compared similar measures with the same standardized test. We used hierarchical linear modeling at each grade level to ascertain the ability of the six measures to model weekly growth trajectories over 13 weeks. All measures produced growth rates that were significant across time, for both kindergarten and first grade, with linear growth observed in all measures.

CHAPTER 1

INTRODUCTION AND REVIEW OF THE LITERATURE

Statement of the Problem

The importance of mathematics understanding has never been more critical as our nation continues to move towards an increasingly multifaceted and technology-driven global community. Citizens today must be connected to the global community in work environments, commerce, and communications. Numerous reports confirm the importance of attaining math proficiency for all students, including those who are just beginning their educational careers (National Mathematics Advisory Panel, 2008; National Council of Teachers of Mathematics [NCTM], 2000; National Research Council, 2001).

Researchers and educators have become increasingly concerned about the low level of mathematics proficiency among American students when examining both national and international comparisons (National Research Council [NRC], 2001). The National Assessment of Educational Progress Mathematics Report Card (NAEP; 2007) indicates that mathematics achievement has improved only slightly in fourth and eighth grades over the last ten years in 23 states, and that a troubling gap remains for students from lower income and minority backgrounds, and those with disabilities. The number of students identified with mathematics disabilities (MD) has increased by over 25 percent since 1990 (National Assessment of Educational Progress [NAEP], 2007; National Mathematics Advisory Panel [NMP], 2008). Researchers are undecided on a specific rationale in order to explain the steady increase in identifying students with MD. A variety of potential causes have been postulated, including more sophisticated identification techniques, increasing populations of students from lower SES which may present an environmental cause, and more recently, inadequate instruction (Bowman, Donovan, & Burns, 2001; Ginsberg & Golbeck, 2004; Lyon et al., 2001).

In an effort to focus attention on the problem of low levels of mathematics proficiency, in 2006, the U. S. Department of Education formed a National Mathematics Advisory Panel tasked with examining the current status of mathematics education and recommending broad-ranging solutions to improve mathematics performance among American students (U.S. Department of Education, 2008). The NMP held a number of symposiums and town hall style meetings conducted with nationally known experts in mathematics, mathematics education, educational psychology, educational policy, and other stakeholders. After extensive consultation with these experts and a comprehensive review of the current research base in mathematics education, the NMP developed an Advisory Report in 2008. In the NMP's mission statement, the panel presented six goals for improving mathematics education. These goals were explained as, a) streamline the mathematics curriculum PreK-8 and emphasize a well-defined set of the most critical domains in early mathematics curriculum, and provide interventions based upon student needs, b) increase the use of rigorous research about how children learn by recognizing the advantages of students having a strong start to their mathematics education, c) recognize mathematically knowledgeable classroom teachers as having a central role in mathematics education, d) inform instructional practices through high-quality research, and e) improve practical assessment in order to monitor student progress and provide teachers with information that can guide instructional practice (summarized from NMP, pg. xiii-xiv).

In particular, two goals of the NMP are aligned with the rationale for the present study. First, the notion of increasing attention on early mathematics intervention is important in order to provide students with foundational mathematics knowledge. This supports the importance of early intervention for students prior to the potential of them having serious mathematics difficulties. An integral tool for measuring the effectiveness of mathematics interventions is at the heart the second goal that serves as a rationale for the present study. The NMP goal that supports the implementation of improved formative assessment practices that guide instructional decision-making at all grade levels is an idea that has been advocated by numerous educational researchers (Clarke, Baker, Smolkowksi, & Chard, 2008; Ginsberg & Golbeck, 2004; Fuchs, et al., 2005). For instance, the National Research Council's *Eager to Learn* report (Bowman, Donovan, & Burns, 2001) calls for assessment of young children to be built into the instructional process. The NAEYC/NCTM joint position paper (National Association for the Education of Young Children and National Council of Teachers of Mathematics, 2002) advocates that the teacher should "… support children's learning by thoughtfully and continually assessing all children's mathematical knowledge, skills, and strategies" (p.4). Other reports also stress the integration of assessment and instruction for younger students (Pellegrino, Chudowsky, & Glaser, 2001). Educators should be able to look to the research literature for viable integrated instructional and assessment practices that address early identification and monitoring of young students' performance in mathematics in order to inform teacher's instructional decisions.

One potential approach to improving young students' mathematics achievement is the implementation of early intervention practices for students who are having difficulties or are at risk of difficulties in mathematics. Preventing academic difficulties through focused early intervention designed to meet the needs of students who struggle is garnering attention in both general and special education research (Clarke et al., 2008). Consistent findings demonstrate that remediating academic problems once they have emerged becomes increasingly difficult the longer the problems are unaddressed. This has led to research efforts to identify critical components that predict which students may be at risk for later academic difficulties (Lyon et al., 2001).

There are a number of key principles that form the foundation of early intervention in mathematics (NMP, 2008; Fuchs et al., 2005; Clarke et al., 2008). First, early intervention requires tools that enable the effective screening of students (Clarke et al.). In early intervention, screening is designed to determine the level of risk a student faces for developing a problem in the future. Second, once the student is identified, the student receives some form of intervention of varying intensity based on the severity of need (Clarke et al.). Third, student response or progress is monitored to determine the effectiveness of the intervention as it is implemented. If the student's progress in not sufficient, it may be necessary to implement changes in instruction (Clarke et al.). The rationale for instructional change is to systematically increase the intensity of instruction until learning progress or growth, as evidenced by the frequent monitoring of student performance, reaches a rate that is considered adequate, with the potential of systematically reducing the level of risk that the student may face for long-term difficulties (Clarke et al.).

One of the most prominent weaknesses found in many formative assessments is in the area of technical adequacy (Clarke et al, 2008; Fuchs et al., 2005). Shinn and

Bamonto (1998) stated that technical adequacy can be defined as those psychometric properties within assessments that provide evidence of reliability and validity of instruments for use in a specific context. This definition has been used by numerous researchers to determine the appropriateness of various assessments (Fuchs, et al.; Clark & Shinn, 2004; Ginsburg & Golbeck, 2004). At present however, while work in early mathematics assessment has escalated in the past few years, particularly in the area of screening measures (Lemke & Foegen, 2009; Clarke et al.; Fuchs et al.), there continues to be a need to determine technically adequate progress measures in early mathematics.

Researchers are beginning to develop conceptual analyses useful for creating mathematics assessments for students in primary elementary grades and to create brief assessment measures that can serve as indicators of mathematics performance (Gersten, Jordan, & Flojo, 2005). Black and Wiliam, (1998) report that principles of assessment should focus on improving learning, and when validated assessments are used for improving learning, these assessments can produce impressive gains in student achievement, especially for struggling students in early mathematics. Geary (2004) detailed the advantages of targeting "skills" proficiency in assessments related to early mathematics development and memory processes. Gersten and Chard (1999) reported on the importance of assessment mechanisms to identify students who are struggling with early mathematical concepts and to monitor their progress through strategic interventions designed to improve their mathematics knowledge.

An example of this type of progress monitoring strategy is a formative assessment model developed using Curriculum-Based Measurement, short, easy to administer assessments that teachers can use to inform instructional decision making (CBM) (Deno, 1985). Clarke and Shinn (2004), Chard, Clarke, Baker, Otterstedt, Braun, and Katz (2005), Lembke and Foegen (2009) have examined the use of timed tasks such as number identification (NI), quantity discrimination (QD), and missing number (MN) with kindergarten and first grade students for screening and progress monitoring. However, more systematic research needs to be conducted with a greater number of students, over a longer period of time, using the most efficient means for teachers to collect data within the constraints of the classroom environment.

The benefits of developing a comprehensive progress-monitoring system in early mathematics include the ability to track student growth and to make instructional adjustments based upon student needs, and to identify students who may be struggling with mathematics at an early stage of their educational career. Subsequently this allows educators an opportunity to improve student performance using data-based decision making and provide a foundation of mathematics knowledge.

Review of Related Literature

In the following review of literature, three major themes as they relate to early mathematics progress monitoring will be discussed. First, the importance of early mathematics proficiency will be reviewed. Second, a detailed examination of the principle of early numeracy (also referred to as number sense) will be discussed. Finally, a review of methods of measuring early mathematics proficiency will be presented, including the use of CBM in monitoring the progress of students in early numeracy that may provide an indication of the growth of primary grade students in early mathematics skills.

The importance of early mathematics proficiency

Early mathematics proficiency is described as the basis for evaluating the mathematics achievement of young school children in the U.S. (U. S. Department of Education, 2007). In an effort to bring increased attention to the importance of improving students' mathematics proficiency, the President created the National Mathematics Advisory Panel (NMP) in April 2006. The panel was charged with the responsibilities of relying on scientific evidence and recommending ways to foster greater knowledge of and improved performance in mathematics among American students. Numerous other researchers and professional organizations responded to the call for increased focus on student achievement in mathematics by releasing reports detailing deficiencies and suggesting methods that could improve the mathematics knowledge of U. S. Students, especially younger students (NMP, 2008; NRC, 2001). To further highlight the importance of early mathematics proficiency, the mathematics that children learn from preschool through the middle grades provides the basic foundation for algebra and more advanced mathematics course work (Baroody& Dowker, 2003). Helping students to learn and think mathematically are made more important due to the demands that society places on its citizens in both the workplace and in the community (NMP, 2008; NCTM, 2002, 2007; NRC, 2001).

Definitions of mathematics proficiency. According to the NMP (2008), the term proficiency in mathematics refers to what students should understand as key concepts; to achieve automaticity as appropriate to grade level, develop flexible, accurate, and automatic execution of basic mathematics operations, and use all of these competencies to solve problems.

In addition, a recent report of the National Research Council (NRC; 2001) described mathematics proficiency as containing five intertwining strands: a) conceptual understanding, b) procedural fluency, c) strategic competence, d) adaptive reasoning, and e) productive disposition. The strands reflect how students learn. The work of the NRC has been incorporated into the description of mathematics standards promoted by the National Council of Teachers of Mathematics (NCTM). Included in the NCTM's standards is a position statement that calls for the implementation of the NRC (2001) strands from the beginning of children's school careers, thereby ensuring the earliest opportunity to promote early mathematics proficiency (NCTM, 2002).

The task of providing every student with the learning structures necessary to become proficient in mathematics at an early age is at the forefront of recent research within many different areas of education research, from mathematics education, and education psychology, to special education (Kilpatrick, 2002). The NRC (2001) report included the need for empirical examination of elements of mathematical knowledge at the earliest stages of formal education, including developing accurate and informative assessment practices. Therefore, before attempting to develop any practical assessment tool to assist teachers in making instructional decisions, it is important to understand the primary domain of mathematics that is presented to primary grade students, early numeracy or "number sense."

Early numeracy

Early numeracy, often referred to as number sense or beginning number sense is described as a child's fluidity and flexibility with numbers; the understanding of what numbers mean, and an ability to perform mental mathematics and make comparisons

(Case, 1998). Students with proficiency in number sense can represent numbers in multiple ways depending on the context of the problem; they can recognize magnitude and can recognize gross mathematical errors in addition and subtraction (Case). Griffin, Case, and Siegler (1994) propose that number sense is often informally acquired prior to beginning school and is a necessary component for learning formal arithmetic in the primary grades.

Definitions of early numeracy or number sense. According to the National Mathematics Advisory Panel Report (2008), number sense can be described as an ability to immediately identify the numerical value associated with small quantities, a facility with basic counting skills, and proficiency in approximating the magnitudes of small numbers of objects and simple numerical operations. The panel recommended that teachers should broaden instruction to include estimation, and that text books should explicitly explain the purposes of number sense.

The National Council for Teachers of Mathematics, the primary professional organization that provides educators with standards of practice in mathematics education, stated that number sense is an understanding about numbers that is drawn from varied meanings of "number" (NCTM, 2000). Students with number sense understand that numbers are representatives of objects, magnitudes, relationships, and other attributes; that numbers can be operated on, compared, and used for communication. The hierarchical and sequential nature of mathematics requires that early elementary students successfully achieve standards within a basic content domain as a prerequisite to accomplishing higher-level standards (NCTM). It is the fundamental knowledge that mathematics, grounded in number sense with its rules and operations, provides students with an inherent awareness of numbers that can be used by the student in flexible ways to solve problems (Gurganus, 2004).

Definition used for present study. While researchers have examined various definitional components of early numeracy, this study defined early numeracy using the NMP (2008) description of early numeracy; "the ability to immediately identify the numerical value associated with small quantities, a facility with basic counting skills, and proficiency in approximating the magnitudes of small numbers of objects and simple numerical operations" (p.19). This definition is preferred due its ability to translate student mathematics performance into observable skills, thereby facilitating practical assessment for determining proficiency of early numeracy skills.

Important components of early numeracy skills. Three important aspects of early numeracy include a) the ability to immediately identify the numerical value associated with small quantities, b) a facility with basic counting skills, and c) proficiency in approximating the magnitudes of small numbers of objects and simple numerical operations (NMP, 2008). Another important aspect of early numeracy lies in the notion that it leads to the automatic use of mathematics information and serves as a key component in solving basic arithmetic computation (Gersten & Chard, 1999).

Developing early numeracy skills through instruction. As children begin primary school many begin to solve problems involving single-digit numbers and quantities without having real objects at their disposal to assist them (Griffin, 2005). This is the point where mental mathematics seems to develop and many use their fingers to keep track of how many items they have counted. At this stage, students begin to exhibit several different forms of the count-on strategy that may take a few years for students to develop into a sophisticated count-on strategy that does not employ fingers (Griffin). This description of the development of early mathematical knowledge is aligned with the theory of instruction as propounded by Bruner (1966), for teaching concepts in mathematics, by providing learning experiences which are appropriate to the level at which a learner can process information. As a result, learners have a high rate of success building competence from a "concrete to abstract" learning continuum for attainment of competencies in mathematics at the elementary grade level (Bruner). As students progress in their understanding, numbers acquire meaning for students when they are able to recognize that each number refers to a specific quantity and that numbers provide a means for describing quantity not only in mathematics class, but in everyday language similar to phonics (Griffin).

Neurological processes of developing early numeracy skills. Butterworth (2005) described the development of arithmetical abilities by reviewing studies that tested the notion of "innate numerosity," where humans are perceived to have number-specific capacities from birth. This is slightly different than the view of the development of general cognitive abilities (i.e. reasoning, short and long-term memory, and spatial processing), often referred to as the "Piagetian view" (Butterworth).

Butterworth's description of numerosity includes the ability to understand one-to-one correspondence. The knowledge that sets of numbers can be manipulated to alter the numerosity, and that the sets may not be visible and may be audible, tactile, or abstract. Lastly, numerosity includes the ability recognize small numbers up to four objects without verbal counting (Butterworth). The conclusion promoted by Butterworth, is that the construction of numeracy goes hand in hand with the development of logic, and that a pre-numerical period corresponds to a pre-logical stage. Butterworth (2005) suggests that logical and arithmetical operations constitute a single system that is psychologically natural and inherent due to development of brain processes as children age.

While definitions differ slightly and debate continues as to the origin of early mathematics abilities and deficits, the impact of early numeracy and its foundational importance to future mathematical competence seems without question (Butterworth; Gersten & Chard, 1999; Griffin, 2005; Gurganus, 2004). While differences are apparent in various researchers' sense of numeracy, most would agree that simply possessing mathematical knowledge is not sufficient in establishing mathematics proficiency. Rather, mathematical proficiency is dependent upon adequate levels of fluency or automaticity in order to possess mathematics proficiency (Butterworth; Gersten & Chard; Griffin; Gurganus; McLeod, 2001; Okamato & Case, 1996).

Definition of automaticity. Automaticity with early numeracy tasks require more than rote memory. For these tasks to become part of a child's long-term memory and, therefore a conditioned response when given a fact, mathematics concepts must be linked in memory processes leading to use of working memory (McLeod, 2001). To further highlight the importance of ensuring students have good number sense, Okamato and Case (1996) stated that children require number sense fluency when using basic number skills in applied scenarios as they solve story problems or number combinations. This is a significant concept, as it suggests early numeracy skills are important early indicators of proficiency in mathematics. Additionally, developing measures that can accurately assess these important early skills can lead to future mathematics success by providing teachers with a tool to make informed program decisions and to monitor instructional effectiveness based upon student's needs (Fuchs, et al., 2007; Lembke & Foegen, 2009).

Linking early numeracy proficiency to future mathematics abilities. As primary grade students work through a hierarchy of early numeracy concepts and skills, connections between the mathematical ideas, connections to standards in other disciplines, and connections to everyday events will facilitate making learning more permanent (McLeod, 2001). Students who fail to achieve competency in early numeracy skills are hard pressed to advance in their mathematical knowledge, and there is empirical support for the relationship of failure to achieve early mathematics competency to underlying deficits in learning (Butterworth, 2005; Geary, 1993; Gersten & Chard, 1999).

Methods of assessing early mathematics proficiency

There are numerous types of assessments that schools and teachers may use to identify and diagnose deficits in mathematics competence at all grade levels. At times it seems as if teachers are inundated with options regarding assessment practices that range from formal (standardized) norm-referenced assessments to informal, criterion-referenced assessments, informal mathematics inventories, interviews, questionnaires, and computer-aided assessments. Each of these various types of assessments requires varied levels of training and materials to implement, as well as differences in psychometric properties. Very often, the psychometric properties associated with assessments are referred to as *technical adequacy*, a general term that describes the levels of reliability and validity that a particular assessment possesses. Yet the field appears to lack a coherent framework to provide educators with assessment practices that will yield the most useful information that can be used to provide instructional supports to students at all levels of ability (Methe et al., 2008). Additionally, with the importance of proficiency of early numeracy skills as a foundation for future mathematics learning, teachers must identify and monitor the progress of those students who may be struggling with early numeracy as early as possible to increase the likelihood of success in later school years. Developing empirically validated assessment practices that inform and guide instruction and can be used to monitor student progress have become a primary focus for educational researchers.

Technical adequacy

Research professionals have been trained to evaluate the technical adequacy of their assessment techniques, and very often published measures are accompanied by an increasingly immense array of reliability and validity data. There are, of course, well established psychometric criteria for judging the technical adequacy of measures (Linn, Baker, & Dunbar, 1991). Key among these is criteria that stem from the fundamental concepts of reliability and validity. Messick (1989) described technical adequacy as being an integration of evaluations and professional judgments of the degree to which "empirical evidence as well as theoretical rationales support the adequacy and appropriateness of inferences and actions based on tests scores or other modes of assessment"(p.13). Shinn, Deno, and Espin (2000) described the basic tenants of technical adequacy as the following; alternate-form reliability which indicates the extent to which results generalize to different item samples, inter-rater reliability indicates the extent to which results generalize across assessors, and criterion validity which assesses whether a test reflects a certain set of abilities. To measure the criterion validity of a test, researchers must calibrate it against a known standard or against itself. Comparing the test with an established measure is known as concurrent criterion validity; testing it over a period of time is known as predictive validity (Shinn et al.). Most standardized assessments and a few formative assessments are developed by analyzing construct validity, where the items on the assessment are statistically analyzed to ascertain their fit to a predetermined construct of the domain(s) that the tests are measuring. The construct is very often determined by following empirical precedence or by utilizing a recognized theoretical framework from which the construct is derived (Messick). Standardized assessments such as the WISC-IV and the TEMA were constructed utilizing an analysis of construct validity (Wechsler, 2008; TEMA-3, 2003). Additionally, as more advanced statistical operations exploratory and confirmatory factor analysis, latent construct validity, and multi-level modeling are employed, assessment developers are gaining a greater understanding of the tenants of technical adequacy of psychometric properties. One possible screening and progress monitoring assessment tool that has received empirical scrutiny regarding technical adequacy is CBM.

Formative assessment practices

In other areas such as literacy, research has documented the importance of identifying early learning difficulties through effective research-based formative assessment practices that lead to improved student achievement and promote future student success (Methe, Hintze, & Floyd, 2008; Fuchs & Fuchs, 2001; Daly, Hintze, & Hamler, 2000; Black & William, 1998; Shinn, 1995; Deno, 1989). School districts throughout the nation have implemented comprehensive kindergarten screening procedures to establish entry-level skills and prepare instructional supports including early intervention practices based upon individual student's needs (Methe et al.). Many of these progress monitoring measures assess areas such as speech, language, early literacy and early numeracy (Howell & Nolet, 1999). Initiating universal screening and subsequent progress monitoring assessment processes may reduce the need for special education intervention later in a student's educational career and could serve as an equalizer in increasing opportunities for student success (Methe et al.)

Given the principles of early intervention, improved practical assessment could be of benefit by providing teachers with insights that can guide teaching (Clarke et al., 2008). Teachers often do not have the means to easily gauge how well their students are performing, or when to make instructional changes in order to meet the instructional needs of students (Baroody & Dowker, 2003). The importance of understanding the various domains of mathematics knowledge, including the informal knowledge that children bring to school cannot be underestimated (Baroody & Dowker; Kilpatrick, Swafford, & Findell, 2001). Frequent formative information can be extremely valuable in planning and guiding instruction for students.

Support for formative assessment practices in mathematics. The NCTM (2000) reported that assessment in mathematics should be more than merely a test at the end of instruction to gauge learning. Rather assessment should become an integral part of the instruction that guides teachers and enhances students' learning. NCTM recommends that teachers should continually gather information about their students and make appropriate decisions about such matters as, instruction, content, pacing, review, and enrichment or remediation for students who may be struggling. NCTM warns that assessment practices that are out of "synchronization" with curriculum and instruction give the wrong signals to all those concerned with education. Any assessment of mathematics learning should first and foremost be anchored in important mathematical content. It should reflect topics and applications that are critical to a full understanding of mathematics as it is used in today's world and in students' later lives, whether in the workplace or in later studies (NCTM).

The National Mathematics Panel Advisory Report. NMP (2008) provided insight on the use of formative assessment in mathematics. They recommend that formative assessments must be of the highest mathematical and *technical quality,* having sound psychometric properties such as reliability and validity. They further recommend that the use of formative assessment—the ongoing monitoring of student learning to inform instruction—is generally considered a hallmark of effective instruction in all disciplines, including mathematics. The report highlighted findings of research that suggests when teachers are provided with appropriate training in data utilization, formative assessment can benefit students at all ability levels by providing data that can inform instructional decisions and target interventions toward specific student needs, including enrichment and remediation for those students who need them (Fuchs et al., 2007; Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, & Hamlett, 1993). Two specific recommendations regarding assessment of mathematics are contained in the 2008 report. One, NMP recommends the frequent use of formative assessment, particularly for students in early elementary grades in order to monitor the progress of students at the beginning of their educational careers. Two, the panel recommended that professional organizations, states, and school districts provide tools for teachers to become informed on specific ways to utilize data gained from assessments, making assessment an integral component of instructional practice in mathematics. The benefits of formative assessment appear most promising for use in classrooms due to its ease of implementation, relevance to curricular content, and the use of data which can inform instructional decisions (Foegen, Jiban, Deno, 2007). Formative assessments carry the endorsement of the NMP (2008). Yet, there are other important methods of assessing student proficiency and identifying those students who may need extra supports in mathematics.

Limitations of formative assessments. The primary limitation of most formative assessment lies in the lack of evidence of adequate technical adequacy. Very often, informal assessments are created by curriculum developers who do not subject their assessments to empirical scrutiny to ascertain the technical adequacy of their assessments. Teachers often develop their own formative assessments based upon what they believe is important content that students should master through classroom instruction (Baroody and Dowker, 2003). While these formative assessment practices may appear to have utility for use in the classroom, there is no empirical evidence to support their utility.

Standardized assessments

Purposes of standardized assessments. One method of assessment used for screening and identification of students with mathematics difficulties is standardized assessment, also known as norm-referenced assessments (Taylor, 2009). The primary characteristics of these types of assessment include being normed using a large population from which comparisons can be made, having a standard set of instructions for the use, administration, scoring, and interpretation of findings, and their empirical evidence into the psychometric attributes that support the use of the assessments. These assessments are often utilized in order to make eligibility decisions, placement decisions, and to form IEP goals. The primary assessment questions that can be ascertained by standardized assessments are frequently related to the classification of mathematics disabilities, or the presence of strengths and weaknesses in broad-based mathematics domains (Taylor). One of the major strengths of standardized assessments is the high level of reliability and validity of these assessments and the use of normative sampling to provide large-scale population norms on which to base comparisons across groups and sub-groups. Scores are often reported as standard scores, percentile ranks, z-scores, t-scores, age and grade equivalent scores, and very often unique scores that represent a combination of component scores from particular tests.

Examples of standardized early mathematics tests. It is important to understand the types of assessments available for use in classrooms, including standardized assessments in order to highlight the differences in information and utility between various types of assessments. One example of a standardized assessment used by teachers is The Test of Early Mathematics Ability—3 (TEMA-3, 2003). The TEMA is a norm-referenced assessment intended to identify the level of mathematics ability for children aged 3-0 through 8-11. According to the authors, the TEMA-3 can be used as a criterion-referenced or diagnostic tool for older students who are having difficulty in mathematics. It is important to note that the TEMA-3 can be used to identify students who are significantly ahead or lagging behind their peers in mathematical thinking, while identifying specific strengths and weaknesses in mathematics content (Ginsberg & Baroody, 2003).

The TEMA-3 is easy to administer and score, and it breaks mathematical concepts into manageable sections for assessment. However, as the test only provides comprehensive scores, the practical uses for teachers are rather limited (Bliss, 2006). The developers of this assessment provide minimal assistance to examinees that need to narrow the focus, with the major criticism being that no subscores are available for the various sub constructs (Bliss). Overall, the TEMA-3 provides useful information and does provide a level of information needed to guide instruction, there is no mechanism for monitoring progress and at a cost of nearly three hundred dollars, it does not appear practical for classroom use (Bliss).

Standardized diagnostic or screening assessments. Identifying students who may be struggling with mathematics concepts and operations is very often the first step in determining who may need early intervention supports in the classroom (Fuchs et al., 2007). An alternative or augmentation to standardized assessments could be found in diagnostic screening assessments that seek to identify students who possess specific learning disabilities in mathematics. This type of assessment seeks to differentiate amongst students who may have mathematics difficulties due to inappropriate teaching, behavioral, and health problems, from those whose poor mathematics achievement cannot be attributed to these problems, but rather seems to be an innate deficit that makes acquiring numeracy skills difficult; dyscalculia (Butterworth, 2003).

An example of a diagnostic screening assessment for dyscalculia can be found in the work of Butterworth (2003). The overarching idea behind the Dyscalculia Screener is that children are normally born with specific capacities for simple numerical tasks; specifically, they are born with an understanding of numeracy (Butterworth). After years of research the developers have selected tests that have been validated to be most effective in discriminating dyscalculia within students. The screener comprises three computer-controlled, item-timed tests, Dot Enumeration, Number Comparison, and Arithmetic Achievement. Since speed of response to numerical questions is the measure of ability in the screener, the developers provide a fourth test that purports to distinguish whether a student is responding slowly, or is simply a slow responder by assessing simple reaction time (Butterworth). This test is administered prior to beginning the three timed assessments.

While the screening assessment described above may provide useful information for the identification of young students with mathematics learning disabilities (MLD), the developers report that the screener is not meant to identify students who may be simply experiencing difficulties with mathematical concepts ,having and it is not meant to measure growth in early numeracy skills for students identified with MLD (Butterworth, 2003). These drawbacks limit the practical utility for classroom teachers who seek to identify and monitor students who may be struggling with early numeracy concepts prior to referral for further evaluation for additional services.

Limitations of using standardized tests. While standardized assessments yield important information in terms of the identification of deficiencies in mathematics proficiency, or the risk for developing these difficulties in early elementary students, the vast majority of these assessments are not practical for teacher use due to their high costs, complex scoring methodologies, and limited direct relevance to instructional objectives (Ysseldyke et al., 2004). One important criterion for effective assessment is teacher utility, often referring to ease of administration and scoring, which is a significant drawback of a large number of standardized assessments due to the specialized training for implementing, administration, scoring, and interpretation of results (Ysseldyke et al.).

An example to highlight the limited nature of information found within many standardized assessments comes from a commonly used IQ assessment, the Wechsler Intelligence Scale for Children –IV (Wechsler, 2008). A typical question, "If you buy 3 dozen pencils at 15 cents a dozen, how much change should you get back from one dollar?" A nine year old student is allowed 45 seconds to solve this problem. Critics point out that this type of questioning doesn't differentiate among the many reasons for answering this question incorrectly, nor does it distinguish between those students who answer the problem automatically within five seconds, and those who take the full 45 seconds to solve it with their fingers (Butterworth, 2003).

Essential comparisons of standardized and formative assessments. Several researchers have reported that the advantages of formal standardized assessments are frequently outweighed by their limited practical value due to the high costs of the tests, formalized training requirements which make the of ease of administration difficult, and a lack of instructional relevance to the curriculum being taught daily in today's classrooms (Graham & Harris, 2006). Standardized assessments such as the TEMA-3, and the Dyscalculia Screener, are impractical for teachers to administer due to the need for specialized training in the principles of the assessments, the administration and scoring of the instrument, as well as the time requirements necessary to complete the assessments make these types of formalized assessments difficult administer to students in a classroom setting (Graham & Harris). The expense of these assessments can be prohibitive for schools to incorporate into their curriculum budgets.

Two comparisons of standardized assessment to formative assessment are important in determining appropriate assessments for use in making instructional decisions. And these two are instructional relevance and the ability to monitor student progress. For instance, the NMP (2008) recommends formative assessment due to its instructional relevance when teachers make data-based educational decisions, and to frequently monitor student progress during interventions. It would therefore seem appropriate to evaluate the ability of standardized assessments to exhibit similar traits. On the other hand, an important reason for such comparison is that standardized assessments undergo rigorous empirical evaluation in order to determine *technical adequacy* of the instruments. NMP recommends that formative assessment be of the highest technical quality, therefore it is important to compare the technical features of formative assessment to that of standardized assessment.

Curriculum-based measurement (CBM)

CBM serves as a potential tool for screening and progress monitoring in that it meets the requirements for teacher utility, sensitivity to instructional effectiveness, sensitivity to frequent monitoring progress of student performance, adaptability for use in identification and intervention, and relevance to the issue of measuring multiple skills contained in acquiring mathematics proficiency (VanDerHeyden, Witt, Naquin, and Noell, 2001; Clarke and Shinn, 2004; Fuchs, Fuchs, Compton, Bryant, Hamlett, and Seethaler, 2007; Lembke, Foegen, Whittaker, & Hampton, 2008).

Background of CBM. CBM is a systematic assessment tool used to monitor students' proficiency and progress in a variety of basic skill areas such as, reading, spelling, mathematics and written expression (Deno, 1985, Deno & Fuchs, 1987). These standard fluency measures are considered to be of utility to educators due to their simplicity in construction and administration, and targeted focus on specific skills. CBM has been referred to as an “academic thermometer” to measure students’ academic health. Much like a thermometer serves as an indicator of physical health, CBM functions as an indicator that allows teachers to make informed instructional decisions regarding the learning needs of students (Shinn & Bamonto, 1998). It is important to note that CBM measures are not instructional methods or interventions in and of themselves, rather they serve as indicators of academic proficiency in a targeted skill (Hosp, Hosp, & Howell, 2007).

CBM has numerous distinctive features that include the assessment of student progress toward long term goals, frequent monitoring and decision-making based upon derived data, and most critical is the technical adequacy of CBM measures that validate the ongoing assessment of student progress and instructional decision-making (Hosp, Hosp, & Howell, 2007; Stecker, Fuchs, & Fuchs, 2005).

Primary uses of CBM. Two of the primary uses of CBM are for screening and progress monitoring (Lembke, Foegen, Whittaker, & Hampton, 2008). Screening of students is an important first step in progress monitoring student performance as the initial step in progress monitoring is to identify those students who are having difficulties (Hosp et al., 2007; Shinn & Bamonto 1998;. & Shinn, 1989). Screening is also useful to determine levels of achievement and to potentially identify students who may be at risk for difficulties, make placement decisions, and to evaluate instructional effectiveness (Lembke et al.) Progress monitoring allows educators to chart student data on a frequent basis, and measuring student progress over time (Lembke et al.). Research in CBM for screening and progress monitoring has been conducted in several curricular disciplines such as reading and written expression, while there is only emerging research occurring on the use of CBM for progress monitoring in early numeracy.

Stages of CBM research. Fuchs (2004) details three stages of research in the area of curriculum-based measurement. In the first stage, measures are examined at one static point in time in order to demonstrate basic technical adequacy. This research has been conducted in the aforementioned studies (Lembke & Foegen, 2009; Lee et al., 2008; Fuchs et al., 2007; Chard et al., 2005). In stage two, researchers should examine the slope (rate of progress) in order to determine if the scores on the measure are associated with improved skill in a specific academic area. Research germane to this stage has been conducted by relatively few researchers (Clarke & Shinn, 2004; Lembke, et al., 2008). In Clarke & Shinn (2005), an examination of only 3 data points throughout an academic year were examined, and Lembke et al. (2008) examined progress monitoring for eight weeks. This scarce data to support progress monitoring in early mathematics provides incentive for the present study. The third stage concerns the instructional utility, where it is determined if the measures can be used by practitioners to improve instructional decision-making and improving student performance (Fuchs). While stage three was not the focus of this study, stage three would be the focus of future research stemming from the present study. Stage three has been examined by others in CBM mathematics at the elementary level (Stecker, Lembke, & Foegen, 2008; Fuchs & Fuchs, 2007).

Technical adequacy in progress monitoring CBM research. Research has been conducted to examine the technical adequacy of using CBM to monitor the progress of students; however, no one has provided a definitive standard for levels of acceptable technical adequacy. Urbino (2004) suggests in general technical adequacy research, correlations should be minimally set at .70 using Pearson Product Moment correlation. Other researchers have set a minimum standard of .80 correlations in order to suggest adequate technical adequacy, although most discussion on standard levels of adequate technical adequacy are discussed in research literature on CBM in literacy measures. Shin, Deno, and Espin (2000) examined appropriate methodologies to ascertain technical adequacy of CBM when used as repeated measures. Three primary characteristics were examined and determined to serve as "practical" determinants for technical adequacy for progress monitoring in CBM. First, alternate-form reliability is required due to the routine use of repeated measures of student performance, so using alternate forms of measures to gather multiple data points it a primary component of progress monitoring using CBM (Shin et al.). It is important to note that there were no specific recommendations pertaining to minimum levels of alternate-form reliability suggested. Second, the ability of the measures to capture growth in the academic constructs being measured, both on the group and individual levels (Shin et al.). Third, the criterion validity of progress monitoring needs to be examined. This was rationalized using the premise that in order to evaluate program effectiveness and make predictions about student outcomes, the progress monitoring measures must relate to other important measures (Shin et al.). For the purposes of the present study, we are extending the work of Lembke, Foegen, Whittaker, & Hampton, (2008), Fuchs, Fuchs, Compton, Bryant, Hamlett, and Seethaler (2007), Clarke and Shinn (2004) and Chard, Clarke, Baker, Otterstedt, Braun, and Katz (2005). These researchers have formed a foundation in their research into progress monitoring using CBM in early numeracy, and several of the measures contained in the present study are similar to those in these studies.

CBM for monitoring progress in early mathematics. The distinctive CBM features described above connect with the NMP (2008) recommendations for assessment of early mathematics. Specifically, the NMP recommendations of ensuring that assessments possess the highest technical quality possible and using formative assessment to monitor student progress in order to inform instruction are tenants of CBM. NCTM (2000) recommendations also connect to CBM principles such as, NCTM's assertion that assessment should be an integral part of instruction with information that enhances enrichment or remediation. According to NCTM, assessments need to be synchronized with curriculum and instruction and anchored in important mathematics content so that teachers of mathematics can provide students with a full mathematics understanding. Foegen, Jiben, and Deno (2007) presented a review of literature on progress monitoring in mathematics. While the largest number of studies of progress monitoring in mathematics occurred in the elementary grades, it was reported that very little research had occurred in early elementary grades mathematics (Foegen et al.). The review concluded with the recommendation of increasing research efforts for progress monitoring in the early elementary grades (Foegen et al.)

Beginning CBM research in math. Early research in mathematics CBM has focused on basic facts as the primary means of assessing student proficiency (Marston, 1989). This presents a dilemma for educators who seek to follow a prevention-approach by identifying and intervening at the earliest possible opportunity to avoid the development future and more significant mathematics difficulties (MD). Most young students have not mastered basic facts in kindergarten and first grade, making the use of basic facts in assessments for these children impractical (Gersten & Chard, 1999). Emerging research has focused on the development of indicators for determining those younger students who may be struggling with early numeracy concepts.

Several studies have sought to identify specific CBM measures that could be used as universal screening and progress monitoring of student performance in early numeracy. While not specifically addressing the issue of early identification of MD per se, these studies instead focused on identifying measures that could be used to determine levels of early numeracy proficiency, and are therefore extremely important as they examined a vital component required for empirically validated identification practices; technical adequacy.

CBM progress monitoring research in early mathematics. A distinguishing characteristic between CBM screening measures and CBM progress monitoring measures is that progress monitoring is primarily focused on frequently monitoring the growth of students who may be having difficulties or may be at-risk for developing learning disabilities in a particular content area or during an intervention period. Therefore it is essential to examine potential measures used to monitor the progress of students over time that are sensitive to change (Hosp, Hosp, & Howell, 2007; Shinn & Bamonto 1998; Stecker, Fuchs, & Fuchs, 2005).

Clarke and Shinn (2004) and Chard, Clarke, Baker, Otterstedt, Braun, and Katz (2005) sought to identify potential screening and progress monitoring for students in kindergarten and first grade. In both studies, measures included CBM oral counting (OC), number identification (NI), quantity discrimination (QD), and missing number

(MN) probes. For students in kindergarten, these researchers reported that each measure possessed adequate alternate form and criterion validity ranging from .50 to .69 with the Number Knowledge Test (Okamato & Case, 1996). For students in first grade, oral counting, quantity discrimination, number identification, and missing number probes established adequate reliability and correlations of .60-.79 with standardized tests serving as a criterion.

Clarke and Shinn (2004) included an examination of the measures sensitivity to growth using screening data. They reported that OC appeared to be the most sensitive followed by NI, QD, and MN. It is interesting to note that while OC revealed the greatest sensitivity to growth, it also had the lowest reliability and validity coefficients, while the other measures had higher correlation coefficients, but were less sensitive to measuring growth (Clarke & Shinn). The limits of this study included its small sample size and the use of three data points to measure growth over time is questionable due to the limited repeated measurement data points over time to support growth modeling.

Higher-level analyses of growth trajectories. Subsequent research in the field has subjected the measures to more sophisticated statistical analyses including logistic regression, and hierarchical linear modeling (HLM), with mixed results (Fuchs et al, 2007; Lembke, Foegen, Whittaker, & Hampton, 2008).

Logistic regression. Fuchs, Fuchs, Compton, Bryant, Hamlett, and Seethaler (2007) examined measures of screening and progress monitoring with 667 first grade students who demonstrated mathematics difficulties and followed them through the end of second grade in mathematics. Five screening measures were administered via paper and pencil assessments upon entry into first grade; fact retrieval, CBM computation, number identification/counting, CBM concepts, and a multiskill measure that contained a percentage of each previously listed problem type. Logistic regression modeling was employed to determine whether the four screening measures would predict which students might possess mathematics difficulties in second grade.

Results indicated that three measures, CBM concepts, CBM computation, and the CBM multi-skill measure yielded a strong model fit .809-.847(AUC: area under the regression curve), a measure of discrimination based upon randomly drawn pairs, or the ability of the screening test to correctly identify second grade students with MD from students without MD. An index of AUC ranges from .5 (chance performance) to 1.0 (perfect performance) with >.90 considered excellent, .80-.90 strong, .70-.80 fair, and <.70 poor (Fuchs, et al.). Alternate form reliability was determined using adjacent data points as the unit of interest. These researchers make an important point regarding the difficulty of ascertaining alternate-form reliability when developing progress monitoring tools.

While the primary goal of progress monitoring research is to measure student growth across time, this is at odds with the requirements of analyzing alternate- form reliability by correlating all measures as student change is expected over time, which should result in a decrease in the relationship of scores between early forms and later forms. Since correlational analysis is a determination of the amount of a relationship between variables, it is necessary to consider the relationship between progress monitoring variables that are relatively close in administration (Fuchs et al.). This research concurs with the findings of the aforementioned studies that the use of CBM screening and progress monitoring can positively influence academic achievement for students who are struggling with early numeracy skills (Fuchs et al.).

HLM analysis of longitudinal data

Finding adequate measures of individual change and valid techniques for research on change are problems that have long perplexed educational researchers (Bryk & Raudenbush, 1987). Research on individual change has been plagued by inadequacies in conceptualization, measurement, and design. Brief reviews of these inadequacies follow.

1. Conceptualization. In any research context, a model of the phenomena under study is an important component for guiding research. Yet in most previous research on individual change, the model of individual growth is rarely addressed explicitly (Bryk & Raudenbush).

2. Measurement. Studies of change typically use tests that are developed to discriminate among individuals at a fixed point in time. Their adequacy for distinguishing the rate of change among individuals is rarely considered during the instrument design process. Further, statistical procedures routinely applied to these instruments, such as standardizing the scores to a common mean and variance over time can effectively eliminate the essence of individual growth (Rogosa et al., 1982). Psychometric procedures are needed that enable assessment of the adequacy of instruments for measuring both status and change (Bryk & Raudenbush).

3. Design. Much of the research on change has been based on data on individual status at two time points, for example, scores on a pretest and a posttest. In general, two time points provide an inadequate basis for studying change (Bryk &

Weisberg, 1977; Rogosa et al.1982). Further, even in instances in which data have been collected on multiple occasions, researchers have typically analyzed the data as a series of separate designs multiple time points. Recent developments in the statistical theory of hierarchical linear models (HLMs), however, now enable an integrated approach for studying the structure of individual growth, examining the reliability of instruments for measuring status and change, and investigating correlates of initial status and change (Bryk & Raudenbush).

Singer & Willett (2003) described a longitudinal analysis of repeated measures using HLM that addresses the issue of growth by examining student data on two levels. In level-1, each individual's observed growth is conceived of as a function of an individual growth trajectory plus random error. This trajectory is determined by a set of individual parameters. This level is considered a *Within-Subject Model or Level 1.* In general, Level 1 linear growth model is modeled by the following general formula:

$$Y_{ti} = \pi_{0i} + \pi_{1i}(WEEKS)_{ti} + e_{ti}$$

where Y_{ti} is the early numeracy score of child *i,* π_{0i} is the initial status of child *i* (week 0), π_{1i} is the growth rate for child *i* over the data collection period and represents the expected change during a fixed period of time, and e_{ti} is the random error at Level 1 that is independently and normally distributed with a mean of 0 and variance σ^2.

In Level 2, often called a *Between-Subjects Model*, is generalized by the assumption that the growth parameters vary across individuals. We formulate a between-subjects model to represent this variation. The equation for the Level 2 linear growth model is modeled by the general formula:

$$\pi_{0i} = \beta_{00} + \beta_{01}(Weeks) + r_{0i}$$

where β_{00} is the mean initial status for child *i* (at week 0), β_{01} is the effect of Weeks on the mean growth rate for child *i* (at week 0), r_{0i} is the random effect of the intercept and slope at Level 2.

Research using HLM for longitudinal studies. Seltzer, Choi, and Thum (2003) describe developments in statistical analyses including the recommendation of HLM that employs growth modeling over time for student achievement. The researchers explained that longitudinal studies are appropriate when interest centers on comparing rates of change in achievement for students participating in one program (intervention) with those assigned to another program or treatment (control group). HLM and other models that present growth trajectories will assist researchers in reporting the presence of growth with a robust ability to resolve the frequently violated assumption of independence of observations (Seltzer et al.)

Beretvas (2004) explained the value of HLM in literacy research in light of federal mandates for showing student improvement in AYP. A detailed description of various models of HLM that include longitudinal analysis was presented. An important conclusion supported the use of HLM in longitudinal analysis due to the ability to include covariates or potential predictors in all levels that serve to control for differences in analysis parameters, such as initial scores, teacher effects, or intervention effects, thereby allowing for a more complete consideration of a variety of mediating factors that make educational multilevel analysis difficult (Beretvas).

Shin, Espin, Deno, and McConnell (2004) demonstrated how to apply HLM to repeated measures of CBM in order to model student growth and the relationship to student and instructional variables. The authors established that HLM has advantages over other statistical models such as logistic regression, analysis of variance (ANOVA) and structural equation modeling (SEM). The reported advantages included flexibility in research design by providing robust analysis in the presence of missing data, uneven data collection waves, and non-linear data, while estimating growth rates and their relations to correlates with superior reliability (Shin et al.) CBM lends itself well to HLM due to the premise of providing multiple data points within a relatively short time frame, the superior validity and reliability of CBM measures, and proven sensitivity to small levels of change over time (Shin et al.).

Lembke and Foegen (2009) examined the technical adequacy of several early numeracy screening measures (NI, QD, MN, and quantity array) for students in kindergarten and first grade (n= >300). Results from their study indicated the strongest reliability coefficients of the mid to high .80s for alternate form and test retest reliability for the NI, QD, and MN measures for both grades. Criterion validity was examined using the Woodcock Mini-Battery of Achievement (MBA, 1994), teacher ratings, Stanford Early School Achievement Test (SESAT, 1996), Test of Early Mathematics Ability—3 (TEMA; Ginsberg & Baroody, 2003). Correlations were generally similar to those of the reliability results for each measure in each grade. Predictive validity was examined by correlating student's fall scores on the early numeracy measures with their scores on the spring administration of the spring criterion measure (Lembke & Foegen). Correlations with teacher ratings ranged between .49** to .70**, and with the TEMA-3 correlations ranged from .34* to.68** (*p< .05, **p<.01). The quantity array measure demonstrated the lowest of any of the measures and was not included in subsequent related research that sought to examine NI, QD, and MN as progress monitoring measures to determine sensitivity to student growth over time (Lembke, Foegen, Whittaker, & Hampton, 2008).

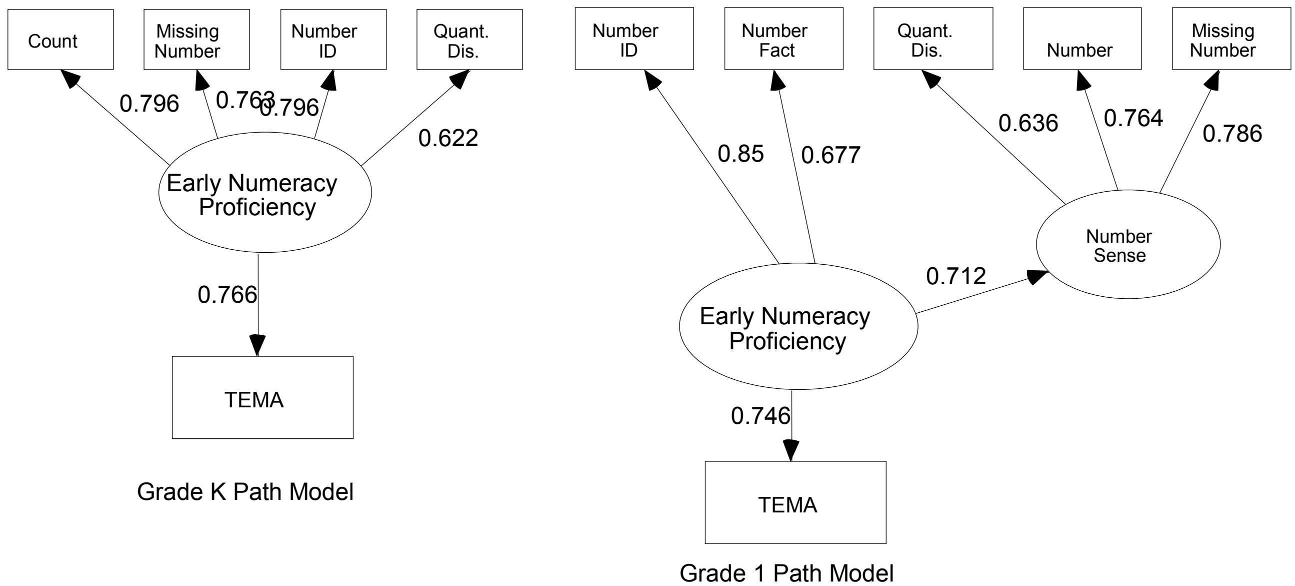
In Lembke et al. (2008), three early numeracy measures were examined as measures of progress monitoring across weeks. NI, QD and MN measures produced growth rates that were significant across time, yet linear growth was indicated for NI only. This suggests some interesting possibilities such as, student growth may be nonlinear due to the nature of skill acquisition during early mathematics instruction, or that MN and QD measures are not as effective as measures of progress (Lembke, et al.). If further studies confirm the curvilinear nature of student performance on MN and QD measures, it could impact the nature of intervention for students who may need additional supports in early mathematics. Educators may need to rethink the decision-making rules to allow additional time for some students to attain these skills (Lembke, et al.).

Research that serves as the foundation of the present study. The six CBM measures that were used in the present study were developed for use in a large-scale study of early numeracy screening measures to assess student proficiency in early mathematics abilities (Lee, Pappas, Lembke, & Ginsberg, 2008). The work of these researchers draws upon two major intellectual traditions. The first is the extensive cognitive science literature on the development of mathematical thinking (Ginsburg, Cannon, Eisenband, & Pappas, 2006; Ginsburg, Klein, & Starkey, 1998). The screening and progress monitoring system includes a focus on the kind of "number sense" shown to be important in predicting academic achievement (Gersten et al., 2005) and on the memory limitations that seem to be an important contributor to mathematics learning problems (Geary, 2004; Jordan, Hanich, & Uberti, 2003; Russell & Ginsburg, 1984).

The second source of theoretical framework is in the methodology of CBM screening and progress monitoring systems. CBM is a simple method of repeated measurement toward long-range instructional goals (Deno, 1985). In order to support the rationale for further study of the CBM measures, investigators collected pilot data with 135 Kindergarten and 1st grade students using the five CBM measures in a suburban school in a small-sized state in New England (Lee et al.). The researchers administered Counting, Missing Number, Number Identification, and Quantity Discrimination measures to the K students, while the 1st grade students were administered Number Identification, Number Facts, Quantity Discrimination, Next Number and Missing Number. All students were also given a standardized mathematics test, the Test of Early Mathematics Ability (TEMA; Ginsburg & Baroody, 2003).

Using the pilot data, the researchers fit several preliminary structural equation models to evaluate the strength of the relationships between ‘Early Numeracy Proficiency’ measured by each of domains of the CBM and the TEMA scores, and the interrelationships among the CBM items. The models selected, presented in Figure 1, were the ones that most closely represent the theoretical relationship underlying the data (Lee et al., 2008).

Figure 1.



The standardized coefficients in the path model for Grade K indicate that a strong relationship exists between 'Early Numeracy Proficiency', and each of variables measured in the CBM battery: Counting (.796), Missing Number (.763), Number Identification (.796), and Quantity Discrimination (.622). 'Early Numeracy Proficiency' also has a strong relationship to the TEMA (.766). All path coefficients were significant at a .05 level. The path model for Grade 1 introduces another latent variable, 'Number Sense', which acts as a mediator between 'Early Numeracy Proficiency' and Quantity Discrimination, Next Number, and Missing Number. The standardized path coefficients show that the strongest relationship exists between 'Early Numeracy Proficiency' and Number Identification (.850). The paths to Quantity Discrimination (.636), Next Number (.764), and Missing Number (.786) were greater using 'Number Sense' as a mediator compared to the basic model, which does not account for any correlations among Missing Numbers, Number Identification, Quantity Discrimination, Next Number, and Number

Facts. Also, the relationship between 'Early Numeracy Proficiency' and the TEMA score turned out to be strong (.746). All path coefficients were significant at a .05 level. The results of the pilot tests of the CBM measures gave the researchers further confidence that these measures could serve as good screening and progress monitoring tools for students in early elementary (Lee et al., 2008). However, further examination with a greater number of students, across several weekly administrations is necessary to validate that the measures are effective for the purposes of progress monitoring in early mathematics.

Statement of Purpose

The present study was conducted to provide an extension of previous research in the technical attributes of six measures used in an early mathematics progress monitoring program. This study represents a Stage Two research effort in the progression of CBM research. The NMP (2008) has presented a proposed agenda for future research into "formative assessment practices" (p.17) that would allow teachers the ability to monitor student progress in their classrooms in order to address student needs. The present research study was in response to the agenda of the NMP by examining a large sample of kindergarten and first grade students over a prolonged study duration of 13 weeks to examine the technical adequacy (reliability, validity, and growth) of specific CBM measures in early numeracy for use in formative assessment. Additionally, this research examined the utility of six measures when monitoring progress over time rather than universal screening at three arbitrary occasions during the academic year. This research has the potential to provide teachers with an assessment practice that would meet the recommendations of the NMP, thereby contributing to improved student outcomes in early mathematics.

Research Questions

1. Are the early mathematics progress monitoring measures technically adequate for use in kindergarten and first grade? (i.e. alternate-form reliability, concurrent criterion validity, predictive validity)
2. Do the progress monitoring measures demonstrate statistically significant growth trajectories for kindergarten and first grade students across 13 weeks?
3. If statistically significant growth trajectories are demonstrated, on which progress monitoring measures do the kindergarten and first grade students exhibit the most growth each week?

CHAPTER 2
METHODS

This study investigated the technical adequacy including determination of weekly growth trajectories of six progress monitoring measures using CBM in early numeracy with kindergarten and first grade students. A correlational design was employed to assess the alternate-form reliability, concurrent criterion validity, and predictive validity of each measure. Additionally, hierarchical linear modeling (HLM) will be used to examine whether there are weekly growth trajectories for each measure over thirteen weeks of weekly progress monitoring, and if the trajectories are linear or curvilinear in nature.

Research study design

The design for this study was an integration of two designs in order to answer the research questions. First, there was a group-design correlational study used to determine the delayed alternate-form reliability and concurrent and predictive validity of the six measures examined in this study. Second, a group-design hierarchical-longitudinal study was utilized in order to examine the 6 measure's ability to capture weekly growth amongst the participants in the study. Both designs worked in tandem so that an overall determination of the measures' adequacy to be used in monitoring the progress of students in kindergarten and first grade could be attained.

Participants and setting

This study took place in a primary elementary school in a small town in a Midwestern state. Support and approval to proceed with the study was granted by the principal and all kindergarten and first grade teachers within the school. All students in the five kindergarten classrooms and the five first grade classrooms received consent forms and those students whose parent or guardians provided consent were selected to be as participants in the study. As an incentive for students to return consent forms from their parents or guardians, students who returned a consent form were given an embossed pencil for returning the consent form to their teachers, whether or not their parents or guardians provided consent to participate in the study.

The participants in the study included 75 first grade students and 71 kindergarten students, for a total sample size of 146 students. The total number of students in K-1 at the school is 206 students, making our sample size 71 percent of the total population of K-1. Of the participants, 70 kindergarten students were Caucasian and 1 was Hispanic. Among the first grade students, all 75 students were Caucasian. This is parallel to the total school population which contains K-2nd grade students, Caucasian students accounting for 95.6 percent of the total population. Of our sample, 17.1 percent qualified for free or reduced price lunch. The special education population for the participants was 4.8 percent and included seven students, with 4 students receiving a classification of speech and language impairment, 2 students with a classification of mild/moderate mental retardation, and 1 student with a classification of attention deficit/hyperactivity disorder. Student demographics are detailed in Table 1. Teacher demographics are contained in Table 2.

Table 1. Demographics of participants

	Male	Female	Caucasian	African-American	Hispanic	IEP	Free and Reduced Lunch
Kindergarten	39	32	70		1	4	11
1st Grade	41	34	75			3	14
Total	80	66	145		1	7	25

Table 2. Demographics of teachers

	Male	Female	Average Years of experience	Master's Degree	Caucasian	Percentage of teachers rated highly qualified
Kindergarten	0	5	18.1	4	5	97.2
First Grade	0	5	12.6	4	5	98.9

Ten data collectors were used during the study and were funded via a combination of an Institutes of Education Sciences grant and university Special Education department funds. The data collectors were comprised of undergraduate and graduate students in education and were appropriately trained in the administration of CBM measures, as well as the mathematics subtests of the WJ-III Tests of Achievement. Training for data collectors consisted of one three-hour session, led by the PI in which an overview of the study was given and an explanation of data collection procedures provided with modeling, guided practice, and independent practice following. Training concluded with completion of fidelity of implementation checklists with each data collector receiving a score greater than .95. No training remediation took place as data collectors each achieved scores higher than the .95 standard set by the PI.

Instrumentation

Each participant completed an alternate form of four (kindergarten) or six (first grade) one-minute early numeracy measures each week for a total of 13 weeks. For kindergarten students, the measures consisted of Counting, Number Identification, Missing Number, and Quantity Discrimination. For the first grade students the measures included, Counting, Number Identification, Missing Number, Next Number, Number

Facts, and Quantity Discrimination. A description of each measure is presented in table 3. The criterion measure for use in determining concurrent criterion validity was the Broad Math Score utilizing mathematics subtests (Calculation, Mathematics Fluency, and Applied Problems) of the Woodcock Johnson Battery of Achievement-III (Woodcock, McGrew, & Mather, 2001). The use of the WJ-III was determined based upon continuing with parallel construction of the present study with an ongoing screening and diagnostic longitudinal study being conducted with the same school and student population.

Table 3. Early Mathematics Measures and Exemplary Items

Counting	Quantity Discrimination	Next Number	Missing Number	Number Identification	Number Facts
Total possible responses: Highest number a student states	60	50	50	80	60
Count as high as you can starting at 1	Randomly presented whole numbers comparisons between 0 and 100, whole numbers to operations, and operations to operations in both addition and subtraction. Alternate the order of placing the larger quantity first or second between sets.	A number is presented orally and the student must state the next number that would follow the number.	3 numbers and a blank are presented in a forward counting sequence by 2's, 3's, 4's, 5's, 6's, and 10'a with the blank varying.	Randomly presented numbers are presented and the student must correctly state the number	Randomly presented addition and subtraction problems with digits less than ten to 2nd graders and addition, subtraction, and multiplication to 3rd graders.
Sample Item	45/20	23_____	2, 3, __	51	5 + 9 = __

Data was collected using a handheld-to-web technology. Specifically, students' answers were imputed on handhelds, which streamlined the administration, scoring, and analysis of each assessment. The handhelds were Palm Pilots that had Mclass Math software installed on each handheld. Wireless Generation, Inc. supplied the handhelds and engineered the Mclass software as part of their in-kind contribution toward the larger, longitudinal study from which the present study was derived. The software system provided the data collector with prompts, which were provided by the examiner. The students responded (either orally or by pointing), and the data collector then used a stylus to record students' responses on the handheld. The student results were then scored automatically and synchronized to a web-based reporting system where the information was restricted with access gained by the PI and dissertation supervisor. The handheld-to-Web system is currently being used in more than 100,000 pre-K-3 classrooms to conduct literacy screening assessments (Sun-Ye et al., 2008).

Assessments that are complex and difficult to administer on paper are simplified because the handheld directs the user to the next task based on the student's response, making it easy for users to focus their attention on students' educational progress and instructional needs, instead of on the assessment administration process (Sharp, 2005). Teachers, who are initially reluctant to use the technology, report that the handheld's ease of use and its clear presentation of the resulting student data overcome their objections (Landry, 2004). The handheld version also introduces considerable efficiency into the assessment process. Evaluations of the literacy assessments show that using handheld assessments can save teachers between 12 and 30 minutes per student, including set-up, administration, data transfer, and analysis, compared to the paper version (Landry, 2004;

Hupert, 2005). Depending on grade, the CBM mathematics assessment items currently take between 4 and 12 minutes to administer per student. While the use of handheld technology has been used in screening in early mathematics (Lee et al., 2008), it has not been applied to progress monitoring in mathematics.

CBM progress monitoring measures

For the Counting task, kindergarten and first grade students were asked to count as high as they could for one minute. For the Number Identification task, both kindergarten and first grade students were asked to identify numbers presented on a page. Kindergarten and first grade students were administered the Missing Number task, which assessed students' ability to identify a missing number from a set of three numbers while counting in a forward fashion with numbers in sequential order. Number sequences included numbers in order, count by five's and count by 10's. The placement of the blank in the number sequence was varied randomly. For the Quantity Discrimination task, kindergarten and first grade students were presented with several types of comparisons and asked to identify the larger quantity. Students were asked to compare whole numbers between 0 and 100, and select either (verbally or by pointing to the number) that represented the larger quantity. The Next Number task was only for first grade students and required them to say the number that comes after a number presented orally by the examiner. The Number Facts task was also only administered to first grade students and assessed students' ability to solve simple addition problems were two addends each consisted of digits less than 10. These problems were presented orally from the examiner and the student responded by stating the answer orally by using "mental mathematics".

Criterion measure

The Broad Mathematics battery of the Woodcock Johnson Tests of Achievement III (WJ-III: Woodcock, McGrew, & Mather, 2001) was used to establish concurrent criterion validity and predictive validity. The Broad Mathematics battery included three subtests: Calculation, Mathematics Fluency, and Applied Problems, with each taking approximately five to ten minutes to administer. The Calculation subtest consisted of addition, subtraction, multiplication, and division problems that the students were presented on a worksheet with problems arranged hierarchically according to the progression of mathematical understanding by age and grade. The Fluency subtest consisted of single-digit addition, subtraction, multiplication, and division problems presented on a worksheet. The student was given three minutes to complete as many of the problems as possible. The Applications subtest consisted of various applied mathematics problems presented to the student using pictures that the examiner flipped from an easel and by reading a prompt that the student orally solved. At a specific point of this subtest, the student could use a pencil and paper to assist them in completing the problem. The WJ III Broad Mathematics battery was appropriate as a measure for criterion validity due to its strong reliability for the mathematics cluster (.95 for ages 5-19), and moderate criterion validity with the Kauffman Test of Educational Achievement (.52) and the Wechsler Individual Achievement test (.70) (Woodcock et al.). The mathematics battery is beneficial for younger students due to its integration of basic counting, facts, and fluency tasks, as well as the ability to discriminate between the proficiency levels of younger students (Woodcock et al.). Additionally, the WJ III has a large standardization sample (8818 individuals across the U.S.), which provides greater ability to generalize across demographic characteristics (Woodcock et al.).

Procedures

The present study took place over 13 weeks, from late January, 2010, through the last week of April, 2010. Each week an alternate but equivalent form of each measure was given to the participants. During the first week, a CBM screening measure designed to be given in the middle of the academic year was administered pre-study and is subsequently called Middle Screening (MS). During the 13th week, a CBM screening measure was given as a post-study measure and was designed to be administered at the end of the academic year, and is subsequently called, End Screening (ES). Each of the progress monitoring measures was administered individually to all students each week. Students were pulled from the classroom and taken to a quiet space in another room. Three measures; Missing Number, Number Identification, and Quantity Discrimination, were presented via a paper student booklet presented by the examiner. Counting, Next Number, and Number Facts was presented orally to each student by the examiner. Fidelity of administration checklists and inter-rater reliability was completed by the PI for all data collectors each week. Table 4 describes the inter-rater reliability attained each week. Data collection was recorded via handheld-to-web technology on a personal digital assistant (PDA). At the completion of the study, the three mathematics subtests of the WJ-III were administered to all participants.

Scoring of the weekly progress monitoring measures consisted of entering the total number of items answered correctly within one minute into the handheld device. All data was uploaded into a web-based data management system which was developed and provided by Wireless Generation, Inc. Broad Math standard scores for the WJ-II were calculated via the WJ-III NU compuscore program. Scoring reliability of the compuscore results were completed by having 40 percent of the tests rescored by another doctoral student who is certified as a school psychologist in the state of Missouri.

Prior to administering the first week's measures, each student was read a "youth assent" by the examiner, with assessments beginning only after gaining affirmative assent by each student. In the event of participant absence, an attempt was made to make-up the assessments during the week. Make-up for a particular week did not occur if that week had passed.

Interrater reliability

Scoring was completed by 10 data collectors under the supervision of the principal student investigator. Interrater reliability was conducted for the CBM measures by having the PI simultaneously entering the student responses along with the data collector for at least 25 percent (ranging from 25 to 27 percent) of administrations each week. Interrater reliability is presented in table 4. Scoring reliability for the WJ-III was completed by having 25 percent of the tests re-scored by a researcher other than the PI, with results reported at 99 percent accuracy. Data entry of scores for each measure into the data sets for analysis were verified by having 40 percent of the total score entries checked by an independent observer other than the PI. The accuracy of data entry was reported at 98.3 percent. Next results associated with each research question will be presented.

Table 4. Interrater reliability

		Week											
Grade	Midscrn	1	2	3	4	5	6	7	8	9	10	11	Endscrn
K	97.3	98.5	99.1	100	100	99.4	97.1	98.6	98.5	96.4	99.1	98.6	99.6
1st	99.2	98.1	98.3	98.2	96.7	97.3	98.5	97.3	99.1	97.1	98.2	97.3	99.1

Independent variables

The CBM measures served as the independent predictor variables, with each participant completing an alternate form of each measure each week for a total of 13 weeks.

Dependent variables

Student performance on the weekly measures and the WJ-III mathematics subtests completed after the 13 week assessment period served as the dependent variable in the study.

CHAPTER III

RESULTS

The purpose of this study was to examine the technical adequacy of progress monitoring using CBM in early numeracy for kindergarten and first grade students. The analyses conducted followed criteria presented in previous studies to identify effective assessment practices; consistent administration and scoring, the ability of the measure(s) to discriminate among students at varied skill levels, concurrent criterion validity, and sensitivity to student growth (Tindal & Parker, 1991). The data analyses for the present study are focused on determining each measure's ability to measure weekly growth, and examining the relationship between student performance on the various measures and student performance on a standardized criterion measure. Analysis was also conducted regarding the potential of the progress monitoring measures to identify growth trajectories over 13 weeks. Descriptive statistics are provided below and results of correlational analyses are reported for research questions one and two. Multilevel modeling was used to determine sensitivity to growth for the progress monitoring measures and will be reported for research question three.

Descriptive analyses

Descriptive statistical analyses were conducted to derive the minimum and maximum number of responses, mean scores, standard deviation, skewness, and kurtosis. In applied statistical analysis, skew and kurtosis denotes observations that are spread on a normal distribution. Kurtosis risk denotes that observations are distributed on the normal curve, thereby determining the independence of the variables. Skew risk denotes that observations are not spread symmetrically around an average value. Waigandt (2004) detailed that skewness plays an important role in hypothesis testing since we must assure that the data set does not violate the assumptions of normality and distribution in order to answer the hypotheses tests. For the present study, this is especially important as one important component of HLM is that we seek to reject the null hypothesis regarding significant differences at both the individual and group levels. Although there are no hard and fast rules regarding minimum and maximum kurtosis and skewness in determining violations of the assumption of normality, it is generally believed that coefficients of less than ± 1.00 provide an adequate indicator of normality (Waigandt). Analyses were completed using data from each week's administration of each measure.

Kindergarten. In the kindergarten Counting measure, students were asked to count as high as possible for one minute. Scores ranged from 11 to 118. Means varied across forms and weeks from 54.15 to 72.44. Standard deviation was 22.58 to 19.62. Skewness and kurtosis coefficients support the assumption of normality each week, with the highest kurtosis occurring on week 5 with a coefficient of -.1.02. Descriptive statistics for kindergarten counting are presented in table 5. There were no ceiling or floor effects noted. Kurtosis on week 2 was reported at -1.018. Visual inspection of a histogram confirmed the assumption of normality.

Table 5. Descriptive statistics: K. Counting

	N	Min	Max	Mean	SD	Skew	Kurt.
Middle Screening Counting	71.00	12.00	109.00	54.15	22.03	.16	-.69
form 1 counting	71.00	16.00	105.00	57.72	21.29	-.07	-.98

form 2 counting	71.00	13.00	97.00	58.21	20.90	-.24	-1.02
form 3 counting	71.00	13.00	105.00	60.32	22.28	-.22	-.85
form 4 counting	71.00	13.00	101.00	59.72	22.58	-.07	-.97
form 5 counting	71.00	12.00	106.00	62.00	20.92	-.25	-.35
form 6 counting	71.00	15.00	109.00	62.65	21.81	-.12	-.66
form 7 counting	71.00	13.00	110.00	65.59	21.38	-.31	-.49
form 8 counting	71.00	16.00	109.00	66.20	22.43	-.27	-.65
form 9 counting	71.00	22.00	109.00	68.79	21.08	-.32	-.73
form 10 counting	71.00	25.00	110.00	72.42	19.97	-.27	-.64
form 11 counting	71.00	25.00	117.00	72.44	19.62	-.28	-.32
end screening counting	71.00	11.00	118.00	72.15	20.83	-.35	-.06

In the Missing Number measure, students were asked to choose the correct number that was missing from a set of three numbers. Scores ranged from 0 to 35. Means varied from 8.79 to 13.42. Standard deviation was from 5.64 to 4.36. Skewness and kurtosis coefficients support the assumption of normality each week. A floor effect was reported on weeks three and six, by one student who was identified with mild/moderate intellectual disabilities. Descriptive statistics for kindergarten missing number are presented in table 6.

Table 6. Descriptive statistics: K. Missing Number

	N	Min	Max	Mean	SD	Skew	Kurt.
Middle Screening MN	71.00	1.00	32.00	8.79	5.34	.60	.45
form 1 MN	70.00	1.00	35.00	9.43	5.59	.77	.32
form 2 MN	71.00	2.00	31.00	10.77	4.81	.86	.35
form 3 MN	70.00	.00	30.00	11.51	5.30	.54	.21
form 4 MN	71.00	2.00	24.00	10.85	4.95	.47	.09
form 5 MN	71.00	2.00	22.00	10.80	4.36	.43	-.20
form 6 MN	71.00	.00	24.00	11.00	5.44	.43	-.45
form 7 MN	71.00	2.00	23.00	11.39	4.90	.12	-.24
form 8 MN	71.00	3.00	26.00	11.38	4.89	.37	.06
form 9 MN	71.00	2.00	26.00	11.89	4.72	.42	.93
form 10 MN	71.00	2.00	27.00	13.03	5.64	.40	-.08
form 11 MN	71.00	4.00	30.00	13.42	5.28	.78	.65
end screening MN	71.00	3.00	30.00	12.87	5.86	.51	.12

In the Number Identification measure, students were asked to name the number shown on a page. Scores ranged from 0 to 70. Means varied from 19.21 to 29.55. Standard deviation was 12.49 to 10.89. Skewness and kurtosis coefficients support the assumption of normality, although kurtosis of 1.080 was reported on the end screening measure. Visual inspection of a histogram confirmed the assumption of normality. A floor effect was reported on form one and form four by two students, one identified with mild/moderate intellectual disabilities, and one student identified with attention deficit/hyperactivity disorder. Descriptive statistics for kindergarten number identification are presented in table 7.

Table 7. Descriptive statistics: Kindergarten Number Identification

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening number ID	71.00	4.00	70.00	20.13	12.47	.67	.42
form 1 number ID	70.00	.00	70.00	19.24	12.35	.53	.45
form 2 Number ID	71.00	4.00	67.00	19.21	11.28	.38	.45
form 3 number ID	70.00	5.00	66.00	21.77	10.89	.51	.51
form 4 number ID	71.00	.00	69.00	22.86	12.01	.35	.21
form 5 number ID	71.00	7.00	63.00	22.83	11.35	.39	.37
form 6 Number ID	71.00	6.00	69.00	22.07	12.04	.67	.54
form 7 number ID	71.00	5.00	70.00	22.68	11.08	.72	.29
Form 8 number ID	71.00	6.00	69.00	23.35	12.49	.45	.64
form 9 number ID	71.00	8.00	70.00	24.13	12.02	.43	.68

form 10 number ID	71.00	8.00	70.00	25.62	12.02	.37	.65
form 11 number ID	71.00	11.00	70.00	27.83	12.49	.30	.46
end screening number ID	71.00	8.00	70.00	29.55	11.86	.23	1.08

In the Quantity Discrimination measure, students are asked to either state or point to the number which represented the larger quantity from a pair of numbers presented on a page. Scores ranged from 0 to 50. Means varied from 16.07 to 24.56. Standard deviation was 9.40 to 6.46. Skewness and kurtosis coefficients support the assumption of normality. A floor effect was reported on the middle screening by one student identified with mild/moderate intellectual disabilities. Descriptive statistics for kindergarten quantity discrimination are presented in table 8.

Table 8. Descriptive statistics: Kindergarten Quantity Discrimination

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening QD	71.00	.00	48.00	16.07	8.08	.64	.37
form 1 QD	70.00	6.00	36.00	17.19	6.46	.63	.01
form 2 QD	71.00	5.00	45.00	21.59	7.53	.44	.84
form 3 QD	70.00	6.00	44.00	19.44	7.73	.38	.30
form 4 QD	71.00	3.00	39.00	20.39	8.57	-.08	-.51
form 5 QD	71.00	2.00	34.00	19.04	6.98	.01	-.32

form 6 QD	71.00	3.00	42.00	20.93	7.77	.32	.17
form 7 QD	71.00	4.00	41.00	22.39	7.82	-.09	-.25
form 8 QD	71.00	6.00	47.00	21.38	8.47	.66	.50
form 9 QD	71.00	2.00	49.00	23.17	9.16	.42	.04
form 10 QD	71.00	2.00	48.00	23.69	9.40	.53	.68
form 11 QD	71.00	8.00	49.00	23.93	8.95	.50	.05
end screening QD	71.00	7.00	50.00	24.56	8.89	.45	.01

First grade. In the first grade Counting measure, students were asked to count as high as possible for one minute. Scores ranged from 29 to 120. Two students achieved the maximum number of 120, indicating a ceiling effect, while no grade 1 students scored 0. Means varied from 78.15 to 91.54. Standard deviation was 17.85 to 16.00. Skewness and kurtosis coefficients support the assumption of normality each week, with the highest kurtosis occurring on week five with a coefficient of 1.03. A visual inspection of a histogram confirmed the assumption of normality. Descriptive statistics for first grade Counting are presented in table 9.

Table 9. Descriptive statistics: 1st grade Counting

	N	Min	Max	Mean	SD	Skew	Kurt.
Middle Screening Counting	75.00	29.00	109.00	78.15	17.32	-.74	.71
form 1 counting	75.00	35.00	109.00	79.88	16.80	-.44	.06

form 2 counting	75.00	39.00	110.00	82.05	16.00	-.60	.11
form 3 counting	75.00	31.00	116.00	85.76	16.58	-.79	.68
form 4 counting	75.00	38.00	120.00	85.33	16.72	-.44	.18
form 5 counting	75.00	29.00	117.00	84.87	16.31	-.91	1.03
form 6 counting	75.00	40.00	119.00	87.71	17.32	-.47	-.02
form 7 counting	75.00	29.00	119.00	87.19	17.85	-.74	.79
form 8 counting	75.00	46.00	120.00	89.89	15.56	-.72	.38
form 9 counting	75.00	48.00	120.00	91.48	16.27	-.44	-.41
form 10 counting	75.00	46.00	120.00	91.43	16.73	-.60	.09
form 11 counting	75.00	46.00	120.00	90.97	17.72	-.58	-.17
End screening counting	75.00	46.00	120.00	91.55	16.27	-.49	-.24

In the Missing Number measure, students were asked to choose the correct number that was missing from a set of three numbers. Scores ranged from 3 to 32. Means varied from 16.60 to 18.37. Standard deviation was from 5.80 to 4.98. Skewness and kurtosis coefficients support the assumption of normality each week. Descriptive statistics for first grade Missing Number are presented in table 10.

Table 10. Descriptive statistics: First grade missing number

	N	Min	Max	Mean	SD	Skew	Kurt.
Middle Screening MN	75.00	3.00	27.00	16.63	5.10	-.24	-.19

form 1 MN	75.00	4.00	28.00	16.64	5.80	-.32	-.44
form 2 MN	75.00	3.00	27.00	16.67	4.98	-.15	-.02
form 3 MN	73.00	5.00	30.00	18.16	5.00	-.23	-.25
form 4 MN	75.00	3.00	29.00	17.32	5.42	-.48	.52
form 5 MN	75.00	4.00	32.00	16.44	5.53	.24	-.18
form 6 MN	75.00	5.00	27.00	16.75	5.07	-.19	-.46
form 7 MN	75.00	3.00	32.00	17.28	5.49	.29	.33
form 8 MN	75.00	7.00	34.00	17.29	5.65	.34	.33
form 9 MN	75.00	7.00	32.00	16.60	5.25	.56	.08
form 10 MN	75.00	7.00	32.00	17.97	5.31	.17	.00
form 11 MN	75.00	6.00	30.00	18.37	5.23	-.08	-.40
end screening MN	75.00	3.00	32.00	17.80	5.71	.12	.40

In the Next Number measure, students are orally presented with a number by the examiner and they must respond with the number that comes after the prompt. Scores ranged from 2 to 35 correct responses. Means varied from 15.78 to 19.68. Standard deviation was from 5.67 to 4.23. Skewness and kurtosis coefficients support the assumption of normality each week, although kurtosis greater than $\pm$ 1.00 was noted on middle screening (1.041), week seven (1.653), and week eight (1.095). A visual inspection of histograms confirmed the assumption of normality. Descriptive statistics for first grade Next Number are presented in table 11.

Table 11. Descriptive statistics: First grade Next Number

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening next number	75.00	5.00	32.00	17.92	4.98	.28	.70
form 1 next number	75.00	4.00	26.00	16.28	4.23	-.81	1.04
form 2 next number	74.00	2.00	24.00	15.78	4.58	-.52	.08
form 3 next number	73.00	6.00	27.00	16.99	4.56	.13	-.38
form 4 next number	75.00	3.00	31.00	16.49	5.67	-.13	-.42
form 5 next number	75.00	4.00	29.00	17.81	5.24	-.10	-.40
form 6 next number	75.00	7.00	27.00	16.68	4.85	-.13	-.49
form 7 next number	75.00	6.00	32.00	17.11	4.63	.68	1.65
form 8 next number	75.00	3.00	34.00	18.49	5.43	.32	1.10
form 9 next number	75.00	7.00	33.00	19.05	5.44	.32	.04
form 10 next number	75.00	2.00	33.00	18.52	5.63	.16	.51
form 11 next number	75.00	7.00	35.00	19.63	5.42	.05	.41

end screening next number	75.00	9.00	31.00	19.68	4.70	-.16	-.23

In the Number Facts measure, students are orally presented with addition problems containing single digit addends. The student must correctly complete "mental mathematics" by orally responding with the correct answer. Scores ranged from 0 to 19 correct responses. On week three and week nine, one student identified with an IEP for mild/moderate intellectual disabilities failed to answer any of the addition problems presented by the administrator. The means varied from 5.09 to 8.24. Standard deviation was from 3.42 to 2.51. Skewness and kurtosis coefficients support the assumption of normality each week, although kurtosis greater than ± 1.00 was noted on form ten (1.53). A visual inspection of histograms confirmed the assumption of normality. Descriptive statistics for first grade Number Facts are presented in table 12.

Table 12. Descriptive statistics: First grade Number Facts

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening number facts	75.00	1.00	12.00	5.12	2.51	.32	.45
form 1 number facts	75.00	1.00	12.00	5.09	2.93	.22	-.40
form2 number facts	75.00	2.00	19.00	6.92	2.88	.91	1.00
form 3 number facts	74.00	.00	17.00	7.69	3.42	.34	.14
form 4 number facts	75.00	1.00	16.00	8.12	3.10	.03	-.09

form 5 number facts	75.00	1.00	11.00	5.59	2.86	.30	-.64
form 6 number facts	75.00	2.00	14.00	7.52	2.68	.23	-.19
form 7 number facts	75.00	3.00	17.00	7.96	2.31	.55	.39
form 8 number facts	75.00	2.00	17.00	7.60	2.79	.45	.66
form 9 number facts	75.00	.00	18.00	7.04	3.41	.64	.43
form 10 number facts	75.00	2.00	19.00	6.81	3.20	.98	1.53
form 11 number facts	75.00	3.00	16.00	8.24	2.81	.69	.59
end screening number facts	75.00	1.00	14.00	7.29	2.57	-.11	-.23

In the Number Identification measure, students were asked to name the number shown on a page. Scores ranged from 6 to 70. Means varied from 37.97 to 45.50. Standard deviation was 13.37 to 11.69. Skewness and kurtosis coefficients support the assumption of normality. Descriptive statistics for kindergarten Number Identification are presented in table 13.

Table 13. Descriptive statistics: First grade NI

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening number ID	75.00	12.00	65.00	39.09	12.74	-.08	-.21
form 1 number ID	75.00	13.00	66.00	38.60	11.69	.13	-.49
form 2 Number ID	75.00	14.00	68.00	37.97	12.52	.23	-.52
form 3 number ID	74.00	15.00	66.00	39.96	11.81	.07	-.57
form 4 number ID	75.00	14.00	70.00	39.91	13.30	.29	-.15
form 5 number ID	75.00	15.00	70.00	40.73	12.44	.18	-.49
form 6 Number ID	75.00	10.00	70.00	40.08	13.37	-.04	-.32
form 7 number ID	75.00	6.00	70.00	41.25	13.28	-.05	-.22
Form 8 number ID	75.00	16.00	70.00	40.76	12.20	.05	-.56
form 9 number ID	75.00	13.00	70.00	41.93	13.18	.20	-.48
form 10 number ID	75.00	16.00	70.00	42.91	12.64	.24	-.15
form 11 number ID	75.00	22.00	70.00	44.72	12.10	.28	-.64

end screening number ID	75.00	22.00	69.00	45.40	11.98	.08	-.71

In the Quantity Discrimination measure, students are asked to either state or point to the number which represented the larger quantity from a pair of numbers presented on a page. Scores ranged from 7 to 60. On weeks four, six, seven, 11, and end screening, one student achieved a ceiling effect for the measure, with a second student achieving a ceiling on week seven. Means varied from 28.73 to 36.25. Standard deviation was 10.94 to 7.53. Skewness and kurtosis coefficients support the assumption of normality. Descriptive statistics for first grade quantity discrimination are presented in table 14.

Table 14. Descriptive statistics: First grade quantity discrimination

	N	Min	Max	Mean	SD	Skew	Kurt.
middle screening QD	75.00	7.00	47.00	29.99	7.53	-.65	.92
form 1 QD	75.00	12.00	46.00	29.51	7.67	-.25	-.41
form 2 QD	75.00	14.00	54.00	34.15	8.44	-.19	-.11
form 3 QD	75.00	12.00	50.00	32.40	9.01	-.32	.63
form 4 QD	75.00	13.00	60.00	33.52	9.67	.26	.32
form 5 QD	75.00	11.00	59.00	28.73	9.14	.73	.43
form 6 QD	75.00	13.00	60.00	35.91	9.52	.26	.04
form 7 QD	75.00	12.00	60.00	34.52	9.54	.41	.81
form 8 QD	75.00	12.00	59.00	33.52	10.36	.07	-.34

form 9 QD	75.00	16.00	59.00	35.85	9.59	.05	-.46
form 10 QD	75.00	10.00	59.00	36.04	10.16	-.11	-.11
form 11 QD	75.00	10.00	60.00	34.87	10.94	-.02	-.45
end screening QD	75.00	11.00	60.00	36.25	9.73	.00	.01

In the Broad Mathematics battery there were three subtests constituted the Broad Math score; Calculation, Mathematics Fluency, and Applied Problems. For kindergarten students, Broad Math standard scores ranged from 26 to 129, and first grade Broad Math standard scores ranged from 72 to 149. The mean for kindergarten students was 96.1, and for first grade students the mean was 105.48, with standard deviations of 16.8 for kindergarten and 13.17 for first grade. Skewness and kurtosis for kindergarten students was -.82 and 2.14 respectively, and for first grade students skewness and kurtosis was .21 and .82 respectively. While skewness and kurtosis for first grade students and skewness for kindergarten students support the assumption of normality, the kurtosis reported for kindergarten students was well above the ±1.00 that is routinely the outer limit for use when testing for the assumption of normality.

Research Question One

Are the early mathematics progress monitoring measures technically adequate for use in kindergarten and first grade? Technical adequacy was determined by analyzing each measure to determine alternate-form reliability, concurrent criterion validity, and predictive validity.

Alternate form reliability. Alternate-form reliability for both kindergarten and first grade was determined by analyzing Pearson Product Moment correlation coefficients calculated by comparing each form to the adjacent week's form. This methodology was employed due to the need to lessen the effects of student change over time that is inherent in progress monitoring research. It is expected that students will change over an extended period of time, thereby providing a potential confound in comparing alternate forms of all weeks in the study. The research base suggests that the use of adjacent weeks is appropriate in considering alternate form reliability in progress monitoring (Shinn & Bamanto, 1998). This method has been used in previous studies for CBM progress monitoring analyses.

Kindergarten alternate-form reliability is presented in table 15. Correlations were reported in the range of moderate to strong for all measures (.66 to .94). The weakest correlations were reported on the Missing Number and Quantity Discrimination measures, .72 and .66 respectively. The strongest correlation was reported on the Number Identification measure .94. All correlation coefficients were found to be statistically significant at an alpha level of $p<.001$.

Table 15. Kindergarten alternate-form reliability

Form	MS&1	1&2	2&3	3&4	4&5	5&6	6&7	7&8	8&9	9&10	10&11	11&ES
Measure												
Counting	0.85	0.86	0.88	0.90	0.88	0.84	0.87	0.87	0.87	0.85	0.85	0.90
Missing Number	0.80	0.72	0.72	0.75	0.78	0.79	0.81	0.84	0.74	0.72	0.76	0.77
Number Identification	0.91	0.90	0.88	0.88	0.87	0.90	0.91	0.91	0.89	0.94	0.91	0.92
Quantity Discrimination	0.74	0.69	0.76	0.72	0.66	0.75	0.72	0.82	0.88	0.86	0.83	0.85
Note: n=71, all p<.001.												

First grade alternate-form reliability is presented in table 16. The correlations ranged from moderate to strong for all measures (.52 to .91). The strongest levels of alternate-form reliability were found in the Counting and Number Identification measures with coefficients of .91. The weakest levels of alternate-form reliability were reported for Next Number and Number Facts measures, .52 and .63 respectively. Both measures are given orally by the administrator and responses are dependent upon the speed in which the administrator states the prompt and the student responds to the prompt. The highest correlations were reported on the Counting and Number Identification measures, .91 for both measures. All correlation coefficients were found to be statistically significant at alpha level of $p<.001$.

Table 16. First grade alternate-form reliability

Form	MS&1	1&2	2&3	3&4	4&5	5&6	6&7	7&8	8&9	9&10	10&11	11&ES
Measure												
Counting	0.83	0.88	0.89	0.88	0.88	0.89	0.90	0.89	0.87	0.91	0.91	0.85
Missing Number	0.74	0.68	0.69	0.77	0.67	0.73	0.66	0.75	0.80	0.78	0.69	0.71
Next Number	0.52	0.67	0.71	0.67	0.75	0.70	0.72	0.77	0.80	0.75	0.66	0.77
Number Facts	0.63	0.64	0.66	0.78	0.67	0.81	0.68	0.68	0.70	0.71	0.67	0.65
Number ID	0.86	0.91	0.89	0.85	0.89	0.85	0.84	0.85	0.84	0.85	0.87	0.88
Quantity Discrimination	0.73	0.71	0.67	0.73	0.63	0.67	0.84	0.81	0.79	0.79	0.81	0.69

Note. N=75, all $p<.001$.

Concurrent criterion validity. Correlations for criterion validity for both kindergarten and first grade were calculated by using the mean of student scores from week 12 and week 13 for each measure, as compared with the criterion variable, the student's Broad Math Standard Score on the Woodcock Johnson Battery of Achievement III which was administered on week 13. Concurrent criterion validity for kindergarten is presented in table 17. Criterion validity coefficients ranged from weak to low-moderate (.26 to .45), with Counting resulting in the strongest correlation of .45, and the weakest correlation being reported in Quantity Discrimination at .26.

Table 17. Concurrent criterion validity-kindergarten

Measure	
Counting	0.45***
Missing Number	0.29**
Number Identification	0.44***
Quantity Discrimination	0.26*

Note: *p<.05, **p<.01, ***p<.001

Concurrent criterion validity for first grade is presented in table 18. The correlations ranged from low to moderate (.39 to .52), with Number Facts showing the strongest correlation, and Missing Number showing the weakest correlation.

Table 18. Concurrent criterion validity-first grade

Measure	
Counting	0.40***
Missing Number	0.39***
Next Number	0.47***
Number Facts	0.52***
Number Identification	0.49***
Quantity Discrimination	0.48***

Note: *** p<.001.

Predictive validity. One feature of technical adequacy is to test the potential measures' ability to predict student performance over time. For the present study the students' Broad Math mean of middle screening and week 1 form for all measures were correlated with the students' standard scores on the WJ-III 13 weeks later. Predictive validity for kindergarten is presented in table 19. Predictive validity coefficients ranged from low-moderate to moderate (.40 to .56), with Counting resulting in the strongest correlation of .56, and the weakest correlation being reported in Quantity Discrimination at .40.

Table 19. Predictive validity-Kindergarten

Measure	
Counting	0.56**
Missing Number	0.47**
Number Identification	0.45**
Quantity Discrimination	0.40**

Note: ** $p<.01$.

Predictive validity for first grade is presented in table 20. Predictive validity correlations were moderate (.44 to .67) with Number Facts resulting in the strongest correlation of .67, and the weakest correlation being reported in Counting at .44. For first graders, all correlations revealed a moderate significant correlation. The Number Facts measure reported the strongest correlation of .67, and Counting reported the weakest correlation of .44.

Table 20. Predictive validity-first grade

Measure	
Counting	0.44***
Missing Number	0.53***
Next Number	0.56***
Number Facts	0.67***
Number Identification	0.49***
Quantity Discrimination	0.46***

Note: ****p* <.001

Research Question Two and Three

Do the progress monitoring measures demonstrate statistically significant growth trajectories for kindergarten and first grade students across 13 weeks? If statistically significant growth trajectories are demonstrated, on which progress monitoring measures do the kindergarten and first grade students exhibit the most growth each week? The most important purpose of the present study was to determine if the early numeracy measures could provide evidence to suggest sensitivity to growth in student performance each week. To accomplish this task, Two-Level Hierarchical Linear Modeling (HLM) was used to examine if significant improvement over time with respect to kindergarten and first grade participants scores on the early numeracy measures: Counting, Missing Number, Number Identification, Quantity Discrimination, Number Facts, and Next Number. HLM software 7.0 Beta (Bryk & Raudenbush, 2010) with full information likelihood estimation was used when the analyses were conducted. Consideration was given to the fact that some of the students' patterns of growth were not perfectly linear. The use of HLM allows for statistical examination of the growth patterns that are not necessarily perfectly linear. Therefore, a 2^{nd} order polynomial model was fit to the data in order to represent the appropriate growth pattern for the various early numeracy measures.

After examining the fit of both the linear and the 2nd order polynomial models to the data, it was established that a linear growth model fit the growth pattern best for both kindergarten and first grade participants for all early numeracy measures. Therefore, the following two-level model was used to examine student scores on each measure across time:

$$\begin{aligned} &Level-1: \\ &Y_{ti} = \pi_{0i} + \pi_{1i}Time + e_{ti} \\ &Level-2: \\ &\pi_{0i} = \beta_{00} + u_{0i} \\ &\pi_{1i} = \beta_{10} + u_{1i} \end{aligned}$$

where Y_{ti} is the observed outcome score for student *i* at time *t*; *Time* is the measurement time minus 13; so that π_{0i} is the intercept or the expected outcome score for student *i* at the final measurement time (the 13th time point); π_{1i} is the slope or the average growth rate in the outcome measure between each measurement time for student *i*; e_{ti} is the error for student *i*; β_{00} is the average intercept or average outcome score for the group of students at the final measurement time; β_{10} is the average slope or average growth rate between each measurement time for the group of students; and u_{0i} and u_{1i} are the random errors for the average intercept and slope, respectively. Unconditional models were run for each measure to estimate the intra-class correlation coefficients (ICCs) which is described as the amount of variability in level two is accounted for in level one. ICCs for variability at Level 2, as well as reliability estimates, are reported in Table 21.

Table 21. Intraclass correlations (ICCs) percentages and reliability coefficients

Kindergarten		
Dependent variable	ICC	Reliability Estimate
Counting	53.10%	.92
Missing Number	51.13%	.94
Number Identification	16.31%	.90
Quantity Discrimination	11.09%	.91

First Grade		
Dependent variable	ICC	Reliability Estimate
Counting	40.15%	.96
Missing Number	38.23%	.94
Number Identification	41.11%	.90
Quantity Discrimination	8.38%	.95
Next Number	16.91%	.93
Number Facts	9.14%	.94

The results for kindergarten students indicated that there was a significant increase in average growth each week. The Counting measure demonstrated the most average weekly growth (1.54 digit increase), followed by Number Identification average weekly growth rate of (.78 digit increase), followed by Quantity Discrimination with an average weekly growth rate of (.60 digit increase), and lastly Missing Number with the lowest average weekly growth rate of (.31 increase). The average intercepts for the each score were statistically significant and there was significant variation among kindergarten participants in terms of the final scores and growth rates. The results of kindergarten HLM are presented in table 22 through 25.

Table 22. Hierarchical linear model for kindergarten student's counting scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	10.86	1.47	71.84***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	1.54	0.16	9.54***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	74.06	69	285.62***
Growth rate, r_{1i}	1.50	70	364.92***
Level-1 error, e_{ti}	65.07		

Note. **p*<.05; **p<.01; ***p<.001

Table 23. Hierarchical linear model for kindergarten student's missing number scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	2.16	0.02	48.08***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	0.31	.05	6.02***

Random Effects:	Variance	df	χ^2
Initial status, r_{0i}	7.34	69	265.52***
Growth rate, r_{1i}	0.15	70	341.56***
Level-1 error, e_{ti}	7.03		

Note. *p<.05; **p<.01; ***p<.001

Table 24. Hierarchical linear model for kindergarten student's number identification scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	5.08	0.68	80.65***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	.73	0.09	8.45***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	21.33	69	310.74***
Growth rate, r_{1i}	0.43	70	399.57***
Level-1 error, e_{ti}	113.7		

Note. *p<.05; **p<.01; ***p<.001

Table 25. Hierarchical linear model for kindergarten student's quantity discrimination scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	4.20	0.68	49.34***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	0.60	0.07	8.30***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	13.46	69	210.75***
Growth rate, r_{1i}	0.27	70	269.83***
Level-1 error, e_{ti}	113.7		

Note. **p*<.05; **p<.01; ***p<.001

The results for first grade students indicated that there was a significant increase in average growth each week. The Counting measure demonstrated the most average weekly growth (1.16 digit increase), followed by Number Identification with an average weekly growth rate of (.53 digit increase), followed by Quantity Discrimination with an average weekly growth rate of (.48 digit increase), followed by the Next Number (.27 increase), Number Facts with an average weekly growth rate of (.18),and lastly Missing Number with the lowest average weekly growth rate of (.10 increase). The average intercepts for the each score were statistically significant with the exception of Number Facts which did not demonstrate significant differences at the initial status of the group. There was significant variation among first grade participants in terms of the final scores and growth rates. The results of first grade HLM are presented in table 26 through 31.

Table 26. Hierarchical linear model for first grade student's counting scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	7.95	1.50	67.12***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	1.16	0.27	4.30***
Random Effects:	**Variance**	**df**	**χ^2**
Initial status, r_{0i}	28.39	73	202.88***
Growth rate, r_{1i}	0.58	73	260.51***
Level-1 error, e_{ti}	41.87		

Note. *p<.05; **p<.01; ***p<.001

Table 27. Hierarchical linear model for first grade student's missing number scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	0.99	0.49	47.01***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	.10	0.04	2.05*
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	5.11	73	189.47***
Growth rate, r_{1i}	0.10	74	243.25***
Level-1 error, e_{ti}	8.30		

Note. **p*<.05; **p<.01; ***p<.001

Table 28. Hierarchical linear model for first grade student's next number scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	1.93	0.56	37.08***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	.27	.04	7.08***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	2.30	73	98.64*
Growth rate, r_{1i}	0.05	74	126.54***
Level-1 error, e_{ti}	12.02		

Note. **p*<.05; **p<.01; ***p<.001

Table 29. Hierarchical linear model for first grade student's number facts scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	1.24	0.37	24.46***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	0.18	0.03	5.93***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	0.81	73	76.26
Growth rate, r_{1i}	.02	74	97.90*
Level-1 error, e_{ti}	9.28		

Note. **p*<.05; **p<.01; ***p<.001

Table 30. Hierarchical linear model for first grade student's number identification scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	3.75	0.84	69.47***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	0.53	0.08	6.55***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	17.38	73	203.83***
Growth rate, r_{1i}	0.35	74	261.07***
Level-1 error, e_{ti}	25.50		

Note. *p<.05; **p<.01; ***p<.001

Table 31. Hierarchical linear model for first grade quantity discrimination scores

Fixed Effects:	Coefficient	se	*t*-ratio
Model for initial status, π_{0i}			
Mean initial status of child, β_{00}	3.36	0.92	44.53***
Model for growth rate, π_{1i}			
Mean initial status of growth, β_{10}	0.48	0.07	6.81***
Random Effects:	**Variance**	**df**	χ^2
Initial status, r_{0i}	10.55	73	136.84***
Growth rate, r_{1i}	10.54	74	175.54***
Level-1 error, e_{ti}	113.7		

Note. *p<.05; **p<.01; ***p<.001

Summary of results. The objectives of the present study were essentially twofold; one, to determine the technical adequacy of the early numeracy measures for use as progress monitoring tools for kindergarten and first grade students; and two, to determine the ability of the measures to demonstrate significant average weekly growth rates. In regards to alternate-form reliability, the kindergarten measures indicated that Number Identification and Counting measures had the strongest coefficients (.94 and .90 respectively). The weakest coefficient was reported in the Missing Number and Quantity Discrimination measure both with a .70. First grade alternate-form reliability analysis indicated that Number Identification and Counting had the strongest coefficients (both .91), with the weakest coefficient demonstrated by Next Number at .52.

Concurrent criterion validity and Predictive Validity demonstrated that all measures showed significant correlations; and that for kindergarten, Counting had the strongest criterion validity correlation of .45 and Quantity Discrimination the weakest with .26. For predictive validity kindergarten indicated that Counting had the strongest correlation of .56 and Quantity Discrimination had the weakest with .40. For first grade, criterion validity, Number Facts had the strongest correlation of .52 and Missing Number had the weakest with .39. For predictive validity, Number Facts had the strongest correlation with .67 and Counting had the lowest with .44.

The major objective of the present study was in regards to determining significant average weekly growth rates for each grade. All measures for both grades with the exception of Missing Number in first grade showed significant linear growth. For both kindergarten and first grade the Counting measure demonstrated the most average weekly growth rate with 1.54 for kindergarten and 1.16 for first grade. Missing Number had the lowest average weekly growth rate for each grade with .31 for kindergarten and .10 for first grade. The results of all of the analyses provide opportunity for further discussion.

CHAPTER IV

DISCUSSION

Overview

The purpose of this study was to examine the technical adequacy of six CBM measures in early numeracy that could serve as tools for teachers to use in monitoring the progress of kindergarten students and first graders. The ultimate goal for creating technically adequate measures lies in the idea that CBM measures serve as "indicators" of overall proficiency in a particular area, in this case early numeracy proficiency. An example to illustrate this concept is that CBM has often been referred to as an "academic thermometer" to measure students' academic health; much like a thermometer serves as an "indicator" of physical health (Shinn & Bamonto, 1998). CBM functions as an indicator that allows teachers to make informed instructional decisions regarding the learning needs of students (Shinn & Bamonto). It is important to note that CBM measures are not instructional methods or interventions, where the goal is to teach students specific content over a specified time, assessing to ensure that concepts taught on a targeted skill or concept have been mastered. CBM indicators are dissimilar in that regard as they serve as "indicators" of overall academic proficiency in a larger discipline, typically content learned over an academic year (Hosp, Hosp, & Howell, 2007).

In this study, we examined delayed alternate-form reliability, concurrent criterion validity, predictive validity, and the ability of the measures to model growth trajectories over time. The statistical analyses were conducted by comparing number correct in one minute for each measure, carried out using alternate but equivalent forms for 13 weeks.

There are only a relative handful of studies that have examined the technical adequacy of early numeracy CBM progress monitoring measures (Chard et al., 2005; Clark and Shinn, 2004; Fuchs et al., 2007; Lembke et al., 2008). The present study aimed to extend the current knowledge base in several important ways by, a) increasing the number of weekly assessments in examining progress monitoring effects for kindergarten and first grade students in early numeracy, b) examining the technical adequacy of measures that have not received prior analysis and, c) applying advanced statistical analysis in an examination of potential growth modeling for the measures.

First, an examination of the descriptive data for each of the six measures revealed that all of the measures resulted in a wide dispersion of participant scores. This is desirable as it suggests that the measures are able to differentiate between students at various ability levels. In kindergarten students and first graders, the largest standard deviations occurred for the Counting measure, followed by the Number Identification measure. These results are similar to those found by Chard et al. (2004) and Clark and Shinn (2004). When examining the data to ascertain floor and/or ceiling effects, for kindergarten students, there were no floor effects noted with the exception of a single student identified with a developmental delay. There was no ceiling effects reported on any of the kindergarten measures. For first graders, there were two weeks (4 & 9) where one student identified with an intellectual disability, received a zero on the Number Facts measure. For the Quantity Discrimination measure, one student attained the maximum number correct prior to the one minute limit on five different occasions and this also occurred once for one other student. The ceiling effects for quantity discrimination suggest that more items should be added to the measure each week to ensure all students have more opportunities to respond than possible in the time allowed. The floor effects that were observed were with students who have been identified for special education services. Finally, the descriptive data was examined to test assumptions of normality. Although there was slight kurtosis noted on two of the 52 assessments for kindergarten students, and slight kurtosis on six of the 78 assessments, the coefficients were generally less than .18 over the <± 1.00, with two ranging between .63-.50 (NN week seven & NF week nine respectively). Given these findings it was appropriate to continue the examination of determining the technical adequacy of each of the measures for both kindergarten and first grade. Next a discussion of results by research question will be provided.

Are the early mathematics progress monitoring measures technically adequate for use in kindergarten and first grade?

Alternate form reliability is an important consideration when determining technical adequacy, as it assures the user that the available forms of a particular progress monitoring measure can reliably measure the same skill as intended. Previous studies had investigated the reliability and validity of various single-skill CBM screening and progress monitoring measures such as those used in the present study (Chard et al., 2005; Clark and Shinn, 2004; Fuchs et al., 2007; Lembke et al., 2008). Results from these earlier studies showed alternate-form reliability, on average, to be between .75 and .95, with the strongest reliability correlations demonstrated for Oral Counting and Number Identification. These results generally echoed those results in terms of reliability with the exception of Quantity Discrimination for both grades, and the Missing Number measure for first grade.

For kindergarten students, the Number Identification measure demonstrated the strongest alternate-form reliability with correlations ranging between .87 and .94 over 13 forms. The Counting measure revealed the coefficients with the same strength in reliability with a range of .84 and .90. Conversely, the Quantity Discrimination measure demonstrated moderate to moderately strong reliability correlations between .69 and .85. There were two forms of QD that had reliability correlations below .70. The MN measure demonstrated moderate to moderately strong coefficients that ranged between .72 and .84. With the exclusion of the two forms of QD that had correlations below .70, all of the other measures for kindergarten fall within the acceptable range for alternate-form reliability (Shinn et al., 2000; Urbina, 2004;).

First grade alternate-form reliability calculation provided data on two new types of measures to be tested, next number and number facts. These two measures had not been previously examined for technical adequacy in progress monitoring. Analyses revealed that the Counting and Number Identification measures demonstrated the strongest reliability with coefficients that ranged between .83 and .91. Quantity Discrimination for first graders revealed unexpected weaker reliability correlations that ranged between .67 and .84, with four forms demonstrating reliability correlations less than .70. The Missing Number measure demonstrated moderate to moderately strong alternate-form reliability ranging from.66 and .80, with four forms reporting correlation coefficients less than .70. Interestingly, the weakest reliability correlations were demonstrated by the two new measures with Number Facts having a range of .63 and .81, with correlations below .70 on four forms. Also, the Next Number measure showed correlations ranging between 52 and .80, with correlations below .70 on eight forms.

Given the weaker and inconsistent reliability correlations for NF and NN measures, it would appear that certain forms of both measures do not meet the minimum standard for demonstrating adequate alternate-form reliability.

For both grades, alternate-form reliability coefficients for the Quantity Discrimination measure were unexpected as in all previous studies noted earlier, this measure obtained stronger levels of reliability. It is possible, that due to the random placement of items and numbers on each measure, there could be items at the beginning of the measure on some weeks that presented difficulty to some participants. This same rationale could explain the weaker alternate-form reliability demonstrated by the Missing Number measure for first graders. This might cause inconsistent performance levels from week to week, which was the case for many participants in this study. It is important to note that the screening measures contained within the Mclass software, Item-Response Theory analyses were conducted on the Middle of Year and End of Year measures, with realignment of items made according to the findings. Similar IRT analysis could provide answers to this question for the progress monitoring measures.

For the Next Number and Number Facts measures in first grade, the weaker levels of alternate-form reliability are cause for some concern. A similar concern regarding item placement exists with these measures. Additionally these two measures require an oral prompt from the examiner and verbal response from the student, which might result in a potential for inconsistency in rates of responses that is dependent upon the examiner as well as the student.

Concurrent criterion validity is another important analysis for determining the technical adequacy of formative assessments (Chard et al., 2005; Clark and Shinn, 2004;

Fuchs et al., 2007; Lembke & Foegen, 2009; Lembke et al., 2008). For the present study, the broad math subtest of the WJ-III was used as the criterion measure. Three subtests comprise the broad math score: computation, fluency, and applications. For kindergarten, the criterion validity coefficients ranged from .45 for Counting, .44 for Number Identification, .29 in missing number, and Quantity Discrimination with .26. For first grade, the coefficients ranged from .52 for Number Facts, .49 for Number Identification, .48 for Quantity Discrimination, .47 for Next Number .40 for Counting, and .39 for Missing Number. Previous studies have reported that measures with coefficients stronger than .40 as possessing adequate criterion validity (Chard et al.; Clark and Shinn ; Fuchs et al.; Lembke & Foegen; Lembke et al.). For the present study this would suggest that for kindergarten, the Counting and Number Identification measure would demonstrate adequate criterion validity, while in first grade, the Missing Number measure is the only measure that would not be considered to possess adequate criterion validity.

The results for concurrent criterion validity were similar to previous studies which used the WJ-III, but it is important to note that most previous studies have used other criterion measures such as the Number Knowledge Test (Okimato & Case, 1993), and TEMA (Ginsberg & Baroody, 1985) for their analyses. One common concern regarding the use of the WJ-III is related to the targeted population of early elementary grade students. The WJ-III has high overall and age-specific validity as a complete battery, within clusters, and within its measures ranging between .86 to .98 (Mather, Wendling, & Woodcock, 2001). Its lowest reliability and validity coefficients are reported in clusters and measures for use in early elementary grades, and are not recommended for use with preschool students (Mather et al.). The use of the WJ broad math score was required due to its use in the larger longitudinal study from which the present study was derived. Given that the measures for use in the present study as well as the longitudinal study were developed by a researcher who also developed the TEMA, it was deemed inappropriate to use the TEMA as the criterion measure in both studies.

Predictive validity coefficients were calculated by using the mean of scores from measures collected in middle screening and week one, and comparing them with the WJ III test which was administered following week 13. Similar concerns regarding the WJ-III were present in the analysis to determine predictive validity. For kindergarten, the predictive validity was strongest with the Counting measure with a coefficient of .56 followed by Missing Number with a coefficient of .47, Number Identification .45, and Quantity Discrimination with the weakest coefficient of .40. For first grade, predictive validity ranged from .44 to .67. The strongest coefficient was reported in the Number Facts measure .67, with the Next Number measure with .56, Missing Number .53, Number Identification .49, Quantity Discrimination .46, and with the weakest coefficient was the Counting measure with .44. The results were similar to previous studies that examined the technical adequacy of progress monitoring measures, with kindergarten measures Counting and Number Identification receiving the strongest coefficients. The introduction of two measures, Next Number and Number Facts that were not included in previous studies had the strongest predictive validity in this study. The coefficients of the remaining first grade measures performed similarly to previous studies.

The Counting, NI and MN measures for both kindergarten and first grade produced similar coefficients as previous studies. The QD measure for both grades failed to perform comparably to other studies in regards of technical adequacy. It is quite possible that the limited construction rules (i.e. randomized placement of numbers and items on a page) for the measures might account for the unexpected weak correlation coefficients reported in this study. Randomization is a common practice in assessment construction, including CBM construction for screening measures (Good & Jefferson, 1998). Other researchers have used a structured construction method (i.e., restricting the percentage of type of items on a page and placement of easier items at the beginning of the first page) for use in developing progress monitoring measures (Chard et al., 2005; Lembke & Foegen, 2009; Lembke et al., 2008). So, a more structured construction method for the measures in the present study may contribute to stronger coefficients for technical adequacy for use in progress monitoring early numeracy for kindergarten and first grade students.

It is interesting to note that while the two oral measures introduced in the first grade measures (Next Number and Number Facts) produced the weakest alternate-form reliability; they produced some of the strongest coefficients in both concurrent criterion and predictive validity. As stated earlier, it is possible that the randomized construction of the measures could have played a role in the weaker alternate-form reliability. It is equally possible that while interrater reliability and fidelity of implementation checklists during the data collection process were strong, the pace of the measure is directly related to a rhythm that the examiner must establish with the student. This phenomenon is quite different from the other measures which provide a written document for the student to use as a prompt and sets a rhythm that is monitored by the examiner and provides a prompt only if the student does not respond within three seconds. The strong correlation for NN and NF in criterion and predictive validity using the WJ-III broad math could be directly related to the appearance of number facts in each of the three mathematics subtests of the WJ-III, as well as the presence of next number items appearing in the mathematics applications subtest of the WJ-III.

Given the weight of evidence from the various analyses completed to assess the technical adequacy of the six measures, technical adequacy for use in progress monitoring for kindergarten and first grade students were similar to previous studies with the exclusion of Quantity Discrimination. In keeping with previous research, the Counting, Number Identification and Missing Number measures performed to a level that suggest they should be considered technically adequate for the purposes of progress monitoring (at least .70 for reliability and .40 for validity). The Quantity Discrimination measure did not meet the test of technical adequacy in the present study, contrary to previous research results and could not be recommended for use based upon these results. The next number and number facts measures that were used in first grade did not meet the standards set by previous research due to weak alternate-form reliability, but did outperform all other measures in concurrent criterion and predictive validity.

While assessing the technical adequacy of the six measures was an important objective of the present study, the major objective of the present study was in determining if the measures were sensitive to demonstrate student growth over time, and if so, which measure(s) produced the most growth each week.

<u>Do the progress monitoring measures exhibit statistically significant growth trajectories for kindergarten and first grade students across 13 weeks? If statistically significant growth trajectories are displayed, on which progress monitoring measures do the kindergarten and first grade students exhibit the most growth each week?</u>

One of the most important questions that researchers seek to answer in Stage Two research (Fuchs, 2004) in CBM is whether the measures can reliably model progress over time, and if there are measures that are more effective than others at measuring weekly growth. Results of the two-level HLM analysis indicated that both kindergarten and first grade students demonstrated significant linear growth on the Counting measure, with estimated mean weekly growth rates of 1.54 correct answers per week for kindergarten and 1.16 correct answers per week for first grade students. Counting represented the measure with the greatest weekly growth rate in the present study. Previous empirical analysis of the Counting measure used either logistic regression or ANOVA; that make direct comparisons not possible, but growth rates in those studies were (.55 digits per week (Clark and Shinn, 2004). They reported that Counting appeared to be the most sensitive, although it also demonstrated the lowest reliability and validity correlations of the four measures. Despite stronger correlation coefficients, the other measures were less sensitive to growth. One potential explanation may be that measures with lower reliability coefficients combined with the presence of higher sensitivity to growth could have larger standard errors of measurement. Caution is called for in this case, as automatically assuming a measure's utility for use in monitoring progress may not be appropriate; lower reliability increases the likelihood of results that are due to some level of error or chance, and therefore not a "true" estimate of growth (Harvel, 2000). These findings support the previous findings by Chard et al. (2005) that suggested considerable student change from fall to spring on the Counting measure and support the use of this measure to monitor student progress. Although three data points are the minimum number required to establish growth, multiple data points are needed to establish mean growth rates with adequate precision (Singer & Willet, 2003; Raudenbush & Bryk, 2002). In the present study, we examined growth rates for each measure across 13 weeks.

Significant weekly growth was demonstrated for Number Identification measures with kindergarten reporting a weekly growth rate of .73 correct answers per week and first grade .53 correct answers per week. These growth rates are greater than had been observed in Lembke and Foegen (2007) for kindergarten (.17) and first grade (.25), and Lembke et al. (2008) for kindergarten (.34) and first grade (.24). These findings support previous findings that suggested considerable weekly growth and support the use of this measure for progress monitoring.

While the Quantity Discrimination measures performed unexpectedly lower than previous research in terms of technical adequacy, the measure did demonstrate significant weekly growth with kindergarten progressing at a rate of .60 correct answers per week and first grade .48 correct answers per week. This measure also demonstrated greater weekly growth rates than observed in Lembke and Foegen (2007) for kindergarten (.49), and (.12) for first grade and Lembke et al. (2008) for kindergarten (.27) and first grade (.12). These finding support previous findings that suggested considerable weekly growth and support the use of this measure for progress monitoring.

The Missing Number measure demonstrated significant weekly growth over the 13 weeks, with kindergarten growing at a rate of.31 correct answers per week and first grade .10 correct answers per week. First grade weekly growth rates are similar to previous studies utilizing HLM analysis (Lembke et al., 2008). These findings support previous findings that suggested significant weekly growth and support the use of this measure for progress monitoring.

Next Number and Number Facts measures were examined for students in first grade, and do not have previous empirical evidence using HLM to compare results. Both Next Number and Number Facts demonstrated significant weekly growth rates with NN .18 correct answers per week and NF .27 correct answers per week. The Next Number measure demonstrated significant growth to support its use for measuring progress. While the Number Facts measure also demonstrated significant weekly growth, a second important component of determining a measure's value as an effective measure for progress monitoring was not met by the measure: ability to differentiate between students at different ability levels.

It is important for any assessment tool to possess the ability to differentiate amongst various levels of proficiency between student populations. All measures examined in the present study except for Number Facts indicated growth rates that varied significantly among students in both kindergarten and first grade. This suggests that all measures with the exclusion of Number Facts serve as an appropriate measure to differentiate students with a particular grade level.

While mean weekly growth rates in the present study were statistically significant and demonstrated greater growth rates than previous studies, we should extend our examination of the measures through the lens of the classroom teacher. We examined the measures for mean weekly growth rates, but with the exception of Counting, all measures reported growth rates of less than one item per week. So as teachers collect and graph this data, they need to be aware of the number of data points they need to collect to be able to establish a reliable slope. For example, when using MN, it would take over three weeks for kindergarteners to grow one item correct, and for first graders, it would take 10 weeks to grow one item. NI would require two weeks for both kindergarteners and first graders to grow one item. However, one must also consider that when these measures are used for progress monitoring, teachers are collecting data for several weeks or an entire school year. So multiplying the growth rate by the total number of weeks can help the teacher create an expected goal.

Limitations of the Study

Some important limitations in the present study should be considered. First, the participants were selected from one school district in a rural town in a Midwestern state. The homogeneous convenience sample of participants used in the study limits the generalizability of the findings to other school environments with diverse student populations. Threats to the internal and external validity of this study's results posed limitations.

Teacher effects posed a potential limitation to the study, as our study did not include an active role pertaining to the implementation of the study. The 10 classroom teachers for our student participants taught their mathematics content according to district curriculum requirements with no input from investigators. This is important to note as an experimental intervention was not implemented to accompany the weekly assessments, therefore we were unable to control the content that was presented in the classroom from week to week during the study. The teachers and administrators were unable to access participant data during the study in an effort to control for teacher effects in which a teacher might see student performance and independently begin specialized instructional support to those students who may not have been performing at a rate similar to peers. Only kindergarten and first grade students who provided parental or guardian consent and gave affirmative youth assent to participate in the study received the weekly measures from the research team. While only consented students participated in the study, each classroom teacher had a handheld device and received training in administration of the measures from which they can choose to conduct weekly assessment to those students who were not participating in the study.

The decision to use the WJ-III Broad Math Score was an additional limitation for our study. While the WJ-III has strong reliability and validity to support its use as a criterion measure, the lowest coefficients are present for those in early elementary grades, particularly in the mathematics subtests. The use of a standardized mathematics battery designed for and validated for use specifically for children in early elementary grades might have improved the validity results in our study. Additionally, the use of a strict randomized construction rule in terms of numeral placement and item placement may have limited the alternate-form reliability results. The possibility of having the most difficult numerals and items at the beginning of a measure and disproportionately included on any particular page, may have inhibited the rate of responses for younger students from one week to the next. The use of bivariate correlation analysis to establish alternate-form reliability is a basic level of analysis and it appropriate for the sample size and number of data points obtained in our study, but the application of Item Response Theory analysis might have provided a better estimate of alternate-form reliability. Coursework and practical application experience was unavailable for the PI in this study. Although these significant limitations were present in our study, the findings presented in this study do hold promise for future practice.

Implications for Practice and assessment

Despite the limitations of our study, there exist significant implications for practice based upon the results of the present study combined with previous results from earlier studies. It appears that the early numeracy measures of Counting, NI, and MN can be used as technically adequate measures for monitoring the progress of kindergarten students. For first graders, it appears that the Counting and NI measures have the greatest utility for progress monitoring.

In terms of the results of HLM analyses to determine mean weekly growth rates, it appears that all measures for both grades are technically adequate indicators of weekly growth for both kindergarten students and first graders, and are technically adequate as measures that have the ability to discriminate between students at various levels of performance abilities (with the exception of NF). It is important to note that based upon our results, the NF measure was not technically adequate to discriminate between ability levels of students in first grade. The Counting measure demonstrated the largest weekly growth rates in both grades, as well as having strong alternate-form reliability for both grades. Therefore, the Counting measure might be the most effective measure for teacher's use in weekly progress monitoring for their kindergarten and first grade students in assessing progress of early numeracy proficiency. NI also performed well in both grades and can be considered effective for assessing students on their early numeracy progress. This is important as it provides teachers with options based upon these results when considering a potential battery of measures to use in their classroom formative assessment practices. While determining which measures produced the greatest average weekly growth is a central goal of stage 2 CBM research, it should be noted that simply having a measure that produces the greatest growth on a week to week basis is not sufficient to diagnose deficits in mathematical concepts. As stated earlier, CBM measures serve as an indicator of mathematical proficiency. Rather, progress monitoring measures should be considered with other diagnostic data to determine if and where targeted students have problems with the conceptual components of early mathematics.

Recommendations for Future Research

The present study provided an important step in the continued empirical examination of CBM measures in monitoring the progress of students in early numeracy, often referred to as Stage Two CBM research. Continued examination of the measures is warranted by implementing revisions to the construction rules for the measures to include a more structured placement of numbers and items on a page, employing a more diverse sampling of participants, and extending the examination to encompass an entire academic year. Future research could be extended to include the examination of creating a multi-skill measures that combine the single-targeted skill items into one measure. Researchers are beginning to examine such “mixed numeracy” measures in screening and progress monitoring in early numeracy (Fuchs et al., 2007; Lembke et al., 2008).

An important aspect of future research would be to move the empirical examination of the measures into Stage Three CBM research. This stage is the most important phase of CBM research as it is in this stage where we examine the effectiveness of the measures as implemented by classroom teachers and their students. The importance of this stage cannot be understated, as it is in this context that the potential benefits of CBM are truly ascertained. Implementation studies in CBM, particularly in mathematics present a fertile ground for future research. Additionally, by testing implementation of CBM in the classroom by teachers, we can see the impact of the measures upon diverse student communities in a real-world context.

With the federal government's introduction of Response to Intervention (RTI) processes in identifying and supporting those students with learning disabilities and those struggling or at-risk of difficulties, a movement in future CBM research should include examination that is tied to research-based interventions that are implemented in each tier of support provided in an RTI model. Assessment practices in an RTI model are simply the "verb" or simply put, the response in RTI, the intervention is the "subject". Any assessment must be validated to be an effective indicator of proficiency in a student's response to a particular intervention, so there are ample future research opportunities to ensure that assessment practices are accurate, effective, and efficient barometers that teachers can rely upon as they seek to provide the most effective instructional support for their students by having accurate information from which to make instructional decisions.

Conclusion

The body of research supporting the use of CBM in early numeracy for monitoring the progress of students is growing (Chard et al., 2005; Clark and Shinn, 2004; Fuchs et al., 2007; Lembke & Foegen, 2009; Lembke et al., 2008). The present study extended this body of work by providing an examination of different measures administered weekly over a longer period of time than previous studies. The findings of this study are encouraging even in the presence of unexpected results and significant limitations. The findings from this study do suggest that there are measures that are technically adequate to be used in early elementary grades when monitoring the progress of students in early numeracy proficiency. The present study also conducted sophisticated growth modeling to analyze each measure's sensitivity to weekly growth changes, and to differentiate between ability levels of students. These results suggested that all of the measures were technically adequate to measure weekly growth for students in the sample, and that all of the measures with the exclusion of Number Facts were technically adequate in their ability to differentiate between various ability levels of kindergarten and first grade students. This study also serves as a base from which examination into applications of CBM progress monitoring research can continue in multiple areas and contexts, including moving toward Stage Three research in CBM.

Proficiency in early mathematics is foundational for future success during a child's educational career. Developing effective formative assessments that can accurately gauge student progress over time hold tremendous promise for students who may be struggling with mathematical concepts. Developing and validating assessments that serve to inform instructional practice thereby improving mathematics performance in early numeracy skills and arithmetic combinations is of great importance to the field. NCLB (2007) require teachers to use instructional evidence-based practices including evidence-based formative assessment practices in their daily instruction. Additionally, and more importantly, professional organizations such as NCTM and CEC, as well as the National Math Panel Advisory Report (2008) advocate that teachers should be using student data to inform their instruction and to make data-based decisions to support their students.

References

Baroody, A.J., & Dowker, A. (2003). *The development of arithmetic concepts and skills: Recent research and theory.* Mahwah, NJ: Lawrence Erlbaum Associates.

Beretvas, S. N. (2004). Using hierarchical linear modeling for literacy research under no child left behind. *Reading Research Quarterly, 39*, 95-99.

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-144.

Bliss, S. (2006). Test of early mathematics ability—third edition: Test review. *Journal of psychoeducational Assessment, 24*, 85-88.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (Eds.). (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

Bruner, J. S. (1996). *Towards a theory of instruction*. Cambridge, MA: Harvard University Press.

Butterworth, B. (2003). The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry, 46*, 3-18.

Butterworth, B. (2003). *Dyscalculia Screener: Ages 6-14 years*. London: Sir Nelson.

Case, R. (1998). *A psychological model of number sense and its development*. Paper Presented at the annual meeting of the American Education Research Association, San Diego. April.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(3), 3-14.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33,* 234-248.

Daly, E. J., Hintze, J. M., & Hamler, (2000). Improving practice by taking steps toward technological improvements in academic intervention in the new millennium. *Psychology in the Schools, 37*, 61-72.

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children.* New York: Guilford Press.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems For databased special education problem solving. *Focus on Exceptional Children, 19,* 1-16.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring in mathematics: A review of literature. *The Journal of Special Education*, 41, 121-139.

Fuchs, L. S., & Fuchs, D. (2008). Best practices in progress monitoring reading and mathematics at the elementary grades. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology V. p.83.* Bethesda, MD: National Association of School Psychologists.

Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*, 311-330.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*, 493-513.

Fuchs, L.S. (2004). The past, present and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-192.

Fuchs, L. S., & Fuchs, D. (2001). Principles for the prevention and intervention of mathematics difficulties. *Learning Disabilities Research and Practice, 16*, 85-95.

Geary, D. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37* (4), 4-14.

Geary, D. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114*, 345-362.

Gersten, R., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical difficulties. *Journal of Special Education, 33*, 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.

Ginsberg, H. P., & Baroody, A. J. (2003). *Test Early Mathematics Ability*, 3rd Ed. Austin, TX: Pro-ed.

Ginsberg, H.P., Cannon, J., Eisenband, J., & Pappas, S. (2006). Mathematical thinking and learning. In K. McCarthey & D. Phillips (Eds.), *Handbook of Early ChildhoodDevelopment*. Hoboken, NJ: Wiley.

Ginsberg, H. P., & Golbeck, S. L. (2004). Thoughts on the future of research on mathematics and science learning education. *Early Childhood Research Quarterly, 19,* 190-200.

Ginsberg, H. P., Klien, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research into practice. In I.E. Sigel & K.A. Renninger (Eds.), *Handbook of Child Psychology*. Hoboken, NJ: Wiley.

Good, R. H., & Jefferson, G. (1998*).* Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced Applications of Curriculum-Based Measurement.* New York: Guilford Press.

Griffin, S. (2005). Fostering the development of whole-number sense: Teaching mathematics in the primary grades. In M. S. Donovan, & J. D. Bransford (Eds.). *How Students Learn Mathematics in the Classroom*. National Research Council, Washington, DC.

Griffin, S. A., Case, R., & Seigler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice.* (pp. 25-49). Cambridge, MA: MIT Press.

Gurganus, S. (2004). Promote number sense. *Intervention in School and Clinic, 40*, 55-58.

Harcourt Assessment. (1996). *Stanford early school achievement test*, 9th Ed. San Antonio, TX: Psychological Association.

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 Press.

Howell, K. W., & Nolet, V. (1999*). Curriculum-Based Evaluation: Teaching and Decision Making.* Belmont, CA: Wadsworth Publishing.

Hubert, N. (2005). *Supporting teachers use of DIBELS data with handheld computers*. Paper presented at the DIBELS Summit September, 2005, Albuquerque, NM.

Jordan, N. C., Hanich, L. B., & Uberti, H. Z. (2003). Mathematical thinking and learning difficulties. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 359-383). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kaminski, R. A., & Good, R. H. (1992). Toward a technology for assessing early literacy skills*. School Psychology Review, 25*, 215-227.

Kilpatrick, J. (2002). Understanding mathematical literacy: The contribution of research. *Educational Studies in Mathematics, 47,* 101-116.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Landerl, K., Bevin, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8-9 year olds. *Cognition*, *93*, 99-125.

Landry, S. (2004). Testing a model of instructional strategies to enhance prekindergarten children's language, literacy, and math skills. *Developmental Psychology, 30*, 201-218.

Lembke, E. S., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first grade students. *Learning Disabilities Research and Practice, 24,* 12-20.

Lembke, E. S., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technical adequate measures of progress in early numeracy. *Assessment for effective intervention, 3* (4), 206-214.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Lyon, G. R., Fletcher, J.M., Shaywitz, S.E., Shaywitz, B.A., Torgeson, J.K., Wood, F.B., et al. (2001). Rethinking learning disabilities. In C.E. Finn, A. J. Rotherham, (Eds.), *Rethinking special education for a new century*. Washington, DC: Thomas B. Fordham Foundation and the Progressive Policy Institute.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it. In M. R. Shinn (ed.), *Curriculum-based measurement: Assessing special children (pp.18-78).* New York: Guilford Press.

Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ III tests of achievement assessment*. New York: John Wiley & Sons, Inc.

Mazzocco, M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia, 58*, 218-253.

McLeod, J. (2001). Building early numeracy skills. In B. Sornson (Ed*.), Preventing early learning failure.* Alexandria, VA: Association for supervision and curriculum development.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed, pp. 13-103). New York: Macmillan.

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review, 37*, 359-373.

National Assessment of Educational Progress. (2007). *The nation's report card: Mathematics 2007.* Washington, DC: U.S. Department of Education.

National Association for the Education of Young Children and National Council of Teachers of Mathematics. (2002). Position statement. Early childhood mathematics: Promoting good beginnings., from http://www.naeyc.org/about/positions/psmath.asp

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Math Panel Advisory Report (2008). U.S. Department of Education: Washington, D. C.

National Research Council (2001). *Adding it up: Helping children learn mathematics*. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Science and Education. Washington, DC: National Academy Press.

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development, 61,* 27-59.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pelligrino, J. W., & Goldman, S. R. (1987). Information processing and elementary mathematics. *Journal of Learning Disabilities, 20*, 23-32.

Russell, R. L., & Ginsburg, H. P. (1984). Cognitive analysis of children's mathematics difficulties. *Cognition and Instruction, 1*, 217-244.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*, 263-286.

Sharp, R. (2005). Task partitioning for multi-core network processors. In *Lecture Notes In Computer Science.* Berlin: Springer.

Shinn J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*, 164-172.

Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: Big Ideas and avoiding confusion. In M. R. Shinn (Ed*.), Advanced applications of curriculum-based measurement.* New York: Guilford Press.

Shin, J., Espin, C. A., Deno, S. L., & McConnell, S. (2004). Use of hierarchical linear modeling and curriculum-based measurement for assessing academic growth and instructional factors for students with learning disabilties. *Asia Pacific Education Review, 5*, 136-148.

Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress-monitoring data to improve instructional decision making. *Preventing School Failure, 52*, 48-58.

Sun-Lee, Y., Pappas, S., Lembke, E.S., & Ginsberg, H.P. (2008). *Development and Analysis of an integrated screening, progress monitoring, and cognitive assessment system for K-3 mathematics: Initial results*. Paper presented at the Annual Meeting of Paper to be presented at the annual meeting of the American Educational Research Association, New York: NY.

Taylor, R.L. (2009). *Assessment of Exceptional Students: Educational and Psychological Procedures. (8th ed.).* Upper Saddle River, NJ: Pearson.

Tindal, G. A., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and Practice, 6,* 211-218.

Urbina, S. (2004). *Essentials of psychological testing.* Hoboken, NJ: John Wiley & Sons, Inc.

U.S. Department of Education (2002). *Twenty-second annual report to Congress on the implementation of the individuals with disabilities education act*. Washington, DC: Government Printing Office.

VanderHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363-382.

Woodcock, R. W., McGrew, R. & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement,* examiner's manual. Itasca, IL: Riverside Publishing.

Woodcock, R.W., McGrew, K.S., & Werder, J.K. (1994*). Woodcock-McGrew-Werder Mini-Battery of Achievement.* Itasca, IL: Riverside Publishing.

Wechsler, R. (2008). *Wechsler Intelligence Scale for Children—IV*. Upper Saddle River, NJ: Pearson.

Ysseldyke, J., Nelson, J. R., Christenson, S., Johnson, D.R., Triezenberg, H., Sharpe, M. Hawes, M. (2004). What we know and what we need to know about the consequences of high-stakes testing for students with disabilities. *Exceptional Children, 71,* 75-94.

APPENDIX

Assessment of Math Skills in Elementary Students

Your child is invited to be in a research study. This study will look at progress monitoring of math skills in elementary students. Your child was selected as a possible participant because he/she is in Kindergarten or First Grade in Southern Boone R1 schools. Please read this form. Ask any questions you have before agreeing to have your child be in the study.

This study is being done by: Dr. Erica Lembke of the University of Missouri and her doctoral student, David Hampton.

Purpose:
The purpose of this study is to find ways for teachers to look at the beginning math skills of students in grades K-1 using handheld technology.

What your child will be asked to do:
If you agree to have your child be in this study, Dr. Lembke or her helpers will score and record your child's math performance. These are the activities:

For students in Kindergarten and 1st grades:

a. Several 1-minute math activities (for a total of 10 to 15 minutes), given individually. Students will do this once each week for 12 weeks.

Mrs. Lembke or one of her helpers will take your child to a quiet place within or near the classroom to complete these activities. Finally, with your permission, we will get your child's academic records, including information on ethnicity, grades, and lunch status. Also, we will get special education status, and results of state and district test scores.

Risks and Benefits of Being in the Study:
There are no risks with the study.The benefits to being in the study are that the university researchers will be able to provide you and the teacher with more detailed information on your child's math skills. The teacher will then be able to plan better math instruction in the future. In addition, this research will benefit teachers who are trying to look at the progress of their students as they develop math skills. Students who return their consent form (whether signed yes or no) will receive a pencil.

Privacy:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify the children who participate. All records will be kept in a locked file. Only the researchers will have access to those records.

Volunteering:

Your decision whether or not to have your child be a part of the study will not affect your current or future relations with Southern Boone R1 Schools or the University of Missouri. If you decide to have your child be a part of the study, you are free to back out at any time without affecting those relationships.

Contacts and Questions: The researcher carrying out this study is Dr. Erica Lembke. You may ask any questions you have by e-mailing (lembkee@missouri.edu) or calling (573-882-0434) Dr. Lembke. If you have any questions or concerns regarding the study and would like to talk to someone other than the researcher(s), contact the Campus Institutional Research Board, 483 Mc Reynolds, Columbia, MO, 65201; telephone (573) 882-9585.

Please read this to your child:

"Dr. Lembke, from the University of Missouri is trying to find ways to help students in math. She would like you to help her by doing some different math activities during the year. If you want to be a part of her study, you need to sign below. This means that you would be willing to complete the math activities and test."

Statement of Consent (please circle yes or no):

YES

I give permission for my child ______________________________ to

son or daughter's name

be a part of University of Missouri research project on math.

______________________________ ______________

Parent or Guardian (signature) Date

______________________________ ______________

Student (signature or initials) Date

NO

I do not want my child ______________________________to be a part of the University

son or daughter's name

of Missouri research project on math.

______________________________ ______________

Parent or Guardian (signature)

VITA

David Hampton grew up in Northeast Connecticut. He received his B.A. in Secondary Education-Social Studies (2003) from Columbia College and his M.A. in Special Education (2007) from the University of Missouri. He completed his PhD. in Special Education, with an emphasis in Learning Disabilities, at the University of Missouri in 2011.

David will be employed as an Assistant Professor in Mild/Moderate Disabilities at Bowling Green State University in Bowling Green, Ohio.