THE LONG-TERM PREDICTIVE VALIDITY OF EARLY MATHEMATICS CURRICULUM-BASED MEASUREMENT

by

STEPHANIE PETRESHOCK BAGLICI

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2008

This manuscript has been read and accepted for the
Graduate Faculty in Educational Psychology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Georgiana S. Tryon, Ph.D.

Date — Chair of Examining Committee

Mary Kopala, Ph.D.

Date — Executive Officer

Robin Codding, Ph.D.

Marian Fish, Ph.D.

Ida Jeltova, Ph.D.

David Rindskopf, Ph.D.

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE LONG-TERM PREDICTIVE VALIDITY OF EARLY MATHEMATICS CURRICULUM-BASED MEASUREMENT

by

Stephanie Petreshock Baglici

Advisor: Dr. Georgiana Tryon (advisor) & Dr. Robin Codding (advisor in absentia)

Recent federal mandates and educational advocates emphasize accountability, early intervention, and responsiveness to intervention using continuous, brief assessments in order to improve student outcomes in mathematics. Curriculum-Based Measurement (CBM) can help meet these directives. Preliminary research suggests that one may use CBM for early identification and intervention with academic problems in mathematics (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Daly, Wright, Kelly, & Martens, 1997; VanDerHeyden et al., 2001; VanDerHeyden et al., 2004; VanDerHeyden, Broussard, & Cooley, 2006). Specifically, reliability and validity data for CBM of early numeracy skills exist. However, there needs to be replication and extension of research in this area. The purpose of this dissertation was to provide information on the sensitivity and long-term predictive validity of early mathematics CBM administered in kindergarten on other direct and teacher-determined measures of mathematics administered in first grade.

The study assessed 61 students from kindergarten to first grade. Students completed a set of experimental early numeracy CBMs (Tests of Early Numeracy [TEN]; Clarke & Shinn, 2002) in kindergarten and first grade. The study included additional criterion measures (i.e., mathematics-CBM [M-CBM], report card grades, teacher ratings of mathematics skills, and discipline referral data) in the first grade assessment. Repeated

measures ANOVAs explored student growth on the TEN measures. Correlation and regression analyses examined the relationships among the experimental measures and between the experimental and criterion measures.

Results indicated significant relationships between kindergarten and first grade TEN performance. In addition, students showed significant growth on all TEN measures from kindergarten to first grade. A measure of kindergarten students' number line skills, missing number, had the most support as a single indicator of first grade outcomes, in particular of first grade computation skills. In addition, it appears that a kindergarten measure of quantity discrimination is an important indicator of first grade teacher-determined outcomes. Preliminary descriptive data showed that students who received no discipline referrals in the first grade generally scored higher on kindergarten TEN than students who received one or more discipline referrals in the first grade.

Acknowledgements

I would like to acknowledge several important people who were instrumental to the completion of this dissertation. First, I must express sincere gratitude to Dr. Robin Codding for her expertise, encouragement, and most of all, for her loyalty to me and this project. I would also like to thank Dr. Georgiana Tryon, for not only serving as the acting committee chair, but also for her guidance in the final phases of this work. I also owe thanks to Dr. Marian Fish, Dr. Ida Jeltova, and Dr. David Rindskopf, for their contributions to this study. In addition, to the faculty of the CUNY Graduate Center/Queens College Integrated School Psychology Program, I am grateful for the knowledge and skills you have shared with me so that I may pursue a career that brings me such great satisfaction.

Without the help of my Educational Psychology classmates, Mitchelle Johnson, Allison Kert, Maria Russo, Seth Sebold, and Linda Sturges, along with the graduate students from the Long Island University School Psychology program who were involved with this project, I could not have tested all of the wonderful children who participated in the study. Thank you, thank you, thank you. I also need to thank Mariya Shiyko for her time and efforts. I am indebted to the special people in the participating schools who helped facilitate this study.

Finally, to my family, I am forever grateful for your patience, love, and support, especially to my husband, Kadir who never once entertained the thought that I could not finish this undertaking.

Table of Contents

Table of Contents

Table of Contents

Table of Contents

List of Tables

List of Tables

List of Figures

CHAPTER I

Introduction

Competency in mathematics is a critical skill that all students should attain in their schooling. Mathematics skills are not only important for academic and occupational success, but also for daily living situations. That is, most jobs require some level of proficiency, and the application of mathematics skills is essential for successful and independent living at home and in the community (Patton, Cronin, Bassett, & Koppel, 1997). Yet, data indicate that many U.S. students are not developing adequate mathematics skills in their elementary and middle years of schooling. According to the National Assessment of Educational Progress (NAEP, 2005), while students' mathematics performance has improved with time; only 36% of fourth and 30% of eighth grade students performed at or above the *proficient* level as of the year 2005 (NAEP).

Clearly then, there is a great need to improve students' mathematics performance in the U.S. Current mandates such as the No Child Left Behind Act of 2001 (NCLB) seek to improve the educational achievement of students in the U.S. and have stressed increased accountability in education, making assessment a critical issue. In addition to an increased emphasis on assessment and accountability, The President's Commission on Excellence in Special Education (2002) also highlighted early identification of academic difficulties and early intervention as essential factors for improving student achievement. Curriculum-Based Measurement (CBM) of mathematics in primary grades (i.e., kindergarten and first grade) is one mode of assessment that may help schools to address accountability requirements as well as assist with early intervention efforts.

As an introduction, this chapter will provide a brief overview of CBM of early mathematics skills and outline the purpose and findings of this study. Chapter two

describes current national mathematics performance, mathematics learning disabilities, the theory of *number sense*, and assessment approaches to determining difficulties in mathematics and measuring student progress. Additionally, chapter two provides a detailed review of the literature on CBM of early math skills and the purpose of the study. Chapter three describes the methodology. Chapter four presents the results. Finally, chapter five provides a discussion of the findings including implications for practice and future research.

Curriculum-Based Measurement (CBM)

CBM is a form of direct assessment of academic skills with several unique characteristics that distinguish it from other types of measurement. Specifically, CBM is brief, simple to administer, matched to curriculum and intervention, and in particular, is sensitive to short-term growth (Bieber & Choi, 2004; Shapiro, 2004). Researchers specifically developed CBM as a tool for progress monitoring, and it has a long history of research support (e.g., Deno, 1992; Fuchs, 2004). The unique characteristics of CBM are in line with the recommendations made by the President's Commission on Excellence in Special Education (2002) to incorporate formative assessment techniques that inform instructional and intervention practices. Moreover, the National Joint Committee on Learning Disabilities (NJCLD, 2005) specifically emphasized the importance of curriculum-based assessment and consistent progress monitoring to effectively identify and provide services to students with learning disabilities. That is, as one of the most widely researched and psychometrically sound types of formative evaluation, CBM can play an important role in Response to Intervention (RTI), a service delivery framework that stresses empirically-supported instruction, early identification, implementation of

intervention with treatment validity, and progress monitoring. This is especially important given that the most recent re-authorization of the Individuals with Disabilities Education Improvement Act (IDEA, 2004) includes RTI as an alternative to achievement-ability discrepancy approaches for making determinations about the eligibility of students for special education services.

Research has supported the use of CBM of reading for academic screening, instructional planning, and progress monitoring (e.g., Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Shinn, 1989, 1998; VanDerHeyden, Witt, Naquin, & Noell, 2001). Relatively less research exists on mathematics-CBM (M-CBM) in general, and with early grades in particular. Three areas comprise M-CBM: (a) computation/operations, (b) applications/problem solving, and (c) early numeracy. Computation, including single skill and multiple skill operations, is the most frequently researched form of M-CBM and demonstrates concurrent validity (Foegen, Jiban, & Deno, 2007; Skiba, Magnuson, Martson, & Erickson, 1986), reliability (Fuchs, Fuchs, & Hamlett, 1990; Tindal, German, & Deno, 1983), and progress monitoring of performance indicators (Fuchs, Fuchs, Hamlett, & Stecker, 1991). While applied mathematics assessments (e.g., word problems, algebra, data analysis, measurement, geometry, and patterns) have less supporting data; research demonstrates that these types of measures are also reliable and valid indicators of student performance on standardized assessments (Fuchs et al., 1994).

Research indicating that early intervention and formative evaluation of student progress represent two variables related to general increases in student achievement (Good & Brophy, 1986) supports the recent emphasis on early literacy and numeracy CBM tools. Early intervention is critical because it may reduce the severity of learning

difficulties and change potentially poor learning outcomes for children (Fuchs & Fuchs, 2001; VanDerHeyden, Broussard, & Cooley, 2006). Clarke and Shinn (2004) noted that in order to maximize the benefits of early intervention, formative evaluation of children's progress is necessary and research associates this evaluation with increases in academic achievement. Additionally, the *Principles and Standards for School Mathematics* (National Council of Teachers of Mathematics [NCTM], 2000) explicitly state that schools should use early assessment in mathematics to inform teaching and to assist with early intervention. In concert with this focus, researchers have made recent efforts to determine the technical properties of CBM for early numeracy skills.

Early Numeracy and CBM

Number sense theory provides the basis of early numeracy CBM, and early numeracy CBM assesses skills such as counting, number naming, number writing, skills associated with number line knowledge, and the ability to make judgments about quantities. Although there is no established definition of number sense, most researchers describe number sense as an understanding of what numbers mean, fluency and flexibility with numbers, the ability to make comparisons, and the ability to perform mental mathematics (e.g., Berch, 2005; Gersten & Chard, 1999; Kalchman, Moss, & Case, 2001). The NCTM (2000) standards for pre-kindergarten through second grade students include understanding numbers, ways of representing numbers, relationships among numbers, and number systems, all of which reflect number sense skills. Therefore, the objective of early numeracy CBM is to identify brief, repeatable measures that are reliable and valid indicators of number sense, and can inform instructional decisions. Reviews of initial research on early numeracy CBM identified several measures as

corresponding to important number sense concepts such as counting, naming and identifying numbers, making comparisons of magnitude, using number lines, and discriminating among numbers and shapes (Foegen, Jiban, & Deno, 2007; Gersten, Clark, & Jordan, 2007; Gersten, Jordan, & Flojo, 2005). However, given the novelty of these studies, it is important to continue to investigate the established measures as well as to identify other measures or forms of measures that may tap into areas of number sense not previously considered.

Preliminary studies (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Daly, Wright, Kelly, & Martens, 1997; VanDerHeyden et al., 2001; VanDerHeyden et al., 2004; VanDerHeyden, Broussard, & Cooley, 2006) demonstrate that an array of early numeracy CBMs possess adequate reliability, validity, and sensitivity at the pre-school through first grade level. As an extension of this important work, VanDerHeyden et al. (2006) have started to explore the long-term predictive validity of these measures across the early years of children's schooling. Their findings suggest that from pre-school age to kindergarten, student performance is moderately to strongly correlated, that students make significant growth, and that there is some evidence supporting the accuracy of pre-school measures for identifying children in need of intervention.

Further evidence of the sensitivity of CBM of early mathematics over multiple school years is important for understanding the utility of such measures for progress monitoring. Despite the fact that research results have identified progress monitoring as a vital aspect of early intervention, Foegen et al. (2007) note that only two studies examined student growth on early numeracy CBM over time and that future research needs to continue to address this issue. In addition, there have not been explorations of

the long-term relationship between CBM of early mathematics skills and other important pertinent school outcomes, such as the acquisition of basic computation skills and outcomes determined by classroom teachers (e.g., teacher ratings on a standardized rating scale and report card grades), along with behavior indicators. Such research would provide additional types of validity evidence relevant to the school and classroom contexts. This is an important area to explore considering, as VanDerHeyden et al. (2006) point out, that part of the rationale for conducting CBM is to garner ecologically-valid assessment data. Finally, research on the long-term predictive validity of early numeracy CBM will hopefully help to identify which measures are the best indicators of later mathematics skills and therefore, the most useful for identifying students who may benefit from early intervention.

Purpose

The purpose of this dissertation was to explore the long-term predictive validity of a specific set of CBM of early mathematics skills collectively termed, the Tests of Early Numeracy (TEN; Clarke & Shinn, 2002). TEN includes the following measures: (a) Oral Counting (OC), (b) Number Identification (NI), (c) Quantity Discrimination (QD), and (d) Missing Number (MN). These measures assess counting skills, number naming skills, the ability to discriminate between which of two quantities is larger, and the ability to identify the missing number among a string of numbers, respectively. Research addresses the technical features of these measures with promising results (Clarke & Shinn, 2004; Chard et al., 2005; Petreshock, Codding, Johnson, Russo, & Schaffer, 2006). This study extends the research on TEN by addressing the previously mentioned gaps in the literature related to the sensitivity and long-term predictive validity of TEN. In addition,

the study explored technical properties of a newly developed measure.

A sample of 61 students participated in four administration sessions from kindergarten to first grade in a longitudinal design. Repeated measures analyses of variance examined the sensitivity of the TEN measures of student performance over time. Correlational statistics assessed relationships among kindergarten and first grade measures. The relationships between four experimental kindergarten TEN measures and four criterion measures: (a) first grade arithmetic skills as measured by mathematics-CBM (M-CBM), (b) teacher ratings of mathematics skills as measured by a standardized rating scale, (c) first grade end of year overall mathematics report card grades, and (d) first grade disciplinary referrals clarified the long-term predictive validity of TEN. I used correlational statistics as well as regression analyses in order to determine whether kindergarten TEN performance predicted first grade outcomes and further, which TEN measures were the best predictors, to evaluate predictive validity results. Finally, I used correlational statistics to evaluate the technical properties of the new measure created for this study.

The dissertation generated two sets of hypotheses. The first set described growth on TEN measures and relationships among kindergarten and first grade measures. Specifically, the first hypothesis within this set was that students would show significant improvement on the TEN measures from kindergarten through first grade. Results confirmed that growth on TEN measures across kindergarten and first grade was statistically significant and not due to chance, providing evidence for the sensitivity of the TEN across multiple school years. The second hypothesis was that kindergarten TEN performances would show significant relationships with first grade TEN performance.

Correlation analyses confirmed the hypothesis of significant relationships among kindergarten and first grade TEN measures.

The second set of hypotheses described the predictive validity of kindergarten TEN performance for first grade outcomes. The first hypothesis in this set was that kindergarten TEN performance would significantly predict first grade computation skills. Results showed that overall, TEN performance accounted for a significant amount of variance in computations skills. The MN measure appears to be the best individual predictors of first grade computation skills. The second hypothesis was that TEN performance would predict teacher ratings on a standardized academic ratings scale. Collectively, TEN measures were significant predictors of teacher ratings. Students' performance on the QD measure along with performance on the OC and MN measures, depending on the assessment point, were the best predictors of teacher ratings. The third hypothesis that the study supported was that kindergarten TEN performance would significantly predict end of year overall report card grades. Specifically, the QD and MN measures predicted end of year overall mathematics report card grades.

The fourth hypothesis in the second set of hypotheses was that TEN performance would predict problematic behavior, as measured by the number of office discipline referrals (ODRs) students received. Due to extremely low rates of ODRs in the sample, I was not able to explore this hypothesis in depth. Although qualitative analysis suggested performance differences among students who received one or more ODRs and those who did not receive any, with students with no ODRS outperforming students with one or more ODRs; these differences were not statistically significant. The final hypothesis was that a new measure, termed Visual Quantity Discrimination (VQD), of students' ability to

discriminate among pictorially represented quantities would correlate with scores on the TEN measures as well as first grade outcomes. The results showed small to large effect sizes for relationships between VQD and the other measures.

CHAPTER II

Literature Review

This chapter provides an overview of the research investigating mathematics assessment and the link with interventions. First, there is a description of national mathematics performance. Second, the chapter presents a brief definition and description of students with math disabilities. Third, there is a presentation of research investigating possible prerequisite mathematics skills, known as *number sense*. Fourth, the chapter describes assessment purposes and approaches for assessing learning difficulties. Fifth, there is a detailed review of the literature on the psychometric properties of CBM of early mathematics skills. Sixth, the chapter discusses research on the relationship between early mathematics performance and contextually relevant variables as well as problematic school behaviors. Seventh, the chapter concludes with the presentation of the rationale for the study and hypotheses to be tested.

Current Math Performance of Students in the United States

As indicated in the previous chapter, national statistics (NAEP, 2005) demonstrate a need for improvement in mathematics competency among students in the U.S. Specifically, more students should be performing in the *proficient* and above range, and fewer students should be identified at the *below basic* level. The percentage of fourth grade students performing at or above the proficient level has increased from 24% in 2000 to 36% in 2005. Similarly, the percentage of eighth graders performing at or above the proficient level has increased from 26% to 30% from 2000 to 2005. Yet, in 2005, 21% of fourth, and 32% of eighth graders continued to perform below basic levels, supporting national efforts to enhance mathematics knowledge and skills in school-age

children. In a study investigating the arithmetic computation of typically achieving students and students with learning disabilities, Cawley, Parmar, Yan, and Miller (1998) also provided evidence that students' mathematics skills are lacking. When the authors examined computational performance of 229 normally achieving 9 to 14 year-olds, they found that just 81% of typically achieving 14-year olds (grade unspecified) had mastered computational addition, while 85% had mastered subtraction, and only 54% had mastered computational multiplication and division. Therefore, it is evident that a substantial number of students have not mastered basic arithmetic computation skills by the age at which most students are preparing to enter high school. These findings represent a sizeable problem, since it is expected that students master basic computational skills by the fourth grade (Shapiro, 2004) and because basic arithmetic skills are so often essential to academic, occupational, and daily living activities (Patton et al., 1997).

There is also evidence that U.S. students are performing below some of their international peers. For example, the 2003 Trends in International Mathematics and Science Study (TIMSS 2003; Gonzales et al., 2004) compared the mathematics and science achievement of students from various countries. Gonzales et al. reported fourth grade comparisons made among students from 25 countries and eighth grade comparisons made among students from 45 countries. The data used to make these comparisons included performance on math and science assessments developed for TIMSS 2003 as well as student, teacher, and principal responses to questionnaires related to students' schooling and learning experiences. Findings from TIMSS 2003 (Gonzales et al.) illustrated that, while U.S. fourth and eighth grade students scored above total international averages in mathematics and science, they continued to perform below a

number of countries. More precisely, fourth grade U.S. students' mathematics average was lower than mathematics averages in seven other countries including Asian countries such as Singapore, Hong Kong, and Japan, as well as European countries including the Netherlands, Latvia, England, and Hungary. With respect to eighth grade math performance, U.S. students performed, on average, below seven countries including some of those who outperformed U.S. fourth graders (Singapore, Hong Kong, Japan, Hungary, and the Netherlands) as well as Korea and Belgium. Although it is encouraging that the eighth grade students showed improvements in overall performance and relative international standing in mathematics; fourth grade students' overall performance remained similar from 1995 to 2003. Additionally, fourth graders' relative international standing in mathematics decreased over this time period.

Findings from another international study, the Program for International Study Assessment (PISA; Lemke et al., 2004), are more troubling. Forty-one countries participated in PISA, a study examining 15 year-olds' abilities in reading, mathematics, and science literacy every three years (Lemke et al.). PISA focused on problems in the context of everyday life in the domains of reading, mathematics, and science as compared to strictly drawing from the academic curriculum (Dossey, McCrone, & O'Sullivan, 2006). The 2003 PISA study focused on math literacy and cross-curricular problem solving. According to Lemke et al., PISA findings showed that U.S. 15 year-olds' average performance in mathematics literacy and problem solving was lower than international averages. More specifically, 23 countries outperformed the U.S. in terms of average combined math literacy scores and 25 countries outperformed the U.S. in terms of average problem solving scores. In addition, PISA utilized six proficiency levels to

interpret results of student performance, where level 1 is the lowest level of proficiency. The percentage of U.S. students in the lowest level of proficiency was greater than the average international percentage for that level; while the percentage of U.S. students in proficiency levels 4 and above was smaller than the international average percentage for those levels (Lemke et al.). In addition to indications that U.S. students may face challenges in using mathematics on a day to day basis, findings from international studies raise apt concern about the future ability of the U.S. to compete in the global marketplace.

Mathematics Learning Disabilities

While U.S. students' mathematics performance suggests that competency in mathematics is a critical issue for our nation as a whole; there are smaller subsets of students for whom deficits in mathematics are also a substantial concern on an individual level. As many as 4% - 7% of students are identified in the United States with some form of a mathematics learning disability ([MLD]; L.S. Fuchs et al., 2005) and as many as 50% of students classified as learning disabled have mathematics related items on their Individual Education Plan (IEP) (Kavale & Reese, 1992). Students with MLD not only perform at lower levels than their non-disabled peers, but also show more limited progress than typically performing students (Cawley et al., 1998), placing them at greater risk for negative academic and occupational outcomes (Mazzocco & Thomson, 2005). Thus, an overview of MLD is also warranted when considering assessment and intervention with students who are not performing to expectations in mathematics.

There is a large range of MLD reported, partly due to varied definitions of MLD and assessment and identification practices. Researchers have used several terms to

describe students who demonstrate deficits in mathematics skills. For example, some researchers have used the term dyscalculia to describe numerical and arithmetical deficits related to brain injury (Hale & Fiorello, 2004). Gersten, Jordan, and Flojo (2005) use the term *mathematics difficulties,* another conceptualization, to describe both students whose mathematics performance falls in the low average (at or below the 35th percentile) and those students who perform "well below average" (p. 294). Mazzocco and Thompson (2005) also acknowledge low achievement as an indicator of MLD. In part, Mazocco and Thompson support the low achievement conceptualization of MLD due to the lack of another unified definition of MLD. Furthermore, Mazzocco and Thompson argue that using a broader definition of MLD, based on low achievement, may provide a greater understanding of MLD because such a definition allows for the study of a wider group of students.

The fourth edition, text revision, of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR; American Psychiatric Association [APA], 2000) provides another definition, based on a discrepancy between ability and achievement. In addition, IDEA provides a specific description of MLD relevant to school psychologists. Both of these sources are similar in that they provide diagnostic decision making guidelines. In addition, the *DSM-IV-TR* also presents information on the prevalence, expected course, and associated features of mathematics disabilities.

According to the *DSM-IV-TR*, Mathematics Disorder (MD) is one Learning Disorder (LD) that may appear in childhood. The essential feature of MD is mathematical ability that falls substantially below that expected for the child's age, intelligence, and age-appropriate education. Furthermore, the disturbance in mathematics must

significantly impair the student's academic achievement or interfere with activities of daily living that require mathematical skills. Finally, if there is a sensory deficit present, the difficulties in mathematical ability must be in excess of those usually associated with the sensory deficit.

The *DSM-IV-TR* states that MD may appear as early as kindergarten or first grade, for example, in the form of counting difficulties or confusion in number concepts. Research supports this notion and suggests that difficulties with mathematics may appear as early as first grade in the form of holding information in short-term memory while counting and poor conceptual understanding of counting (Geary, 2004). However, students can perform inconsistently on achievement testing across successive academic years, complicating understandings of the course of mathematics difficulties and whether or when a diagnosis of MD is warranted (Geary, 2004). In cases where MD exists, outcomes can be severe. Several sources state that students with LD are at greater risk for dropping out of school and if difficulties persist into adulthood, problems with employment and social adjustment may also surface (*DSM-IV-TR*; Mazzoco & Thompson, 2005).

In the school setting, psychologists make determinations about children's learning problems and eligibility for special education services using criteria established in IDEA (2004). Students with difficulties in mathematics may qualify for special education services under the Specific Learning Disability (SLD) category. IDEA defines SLD:

> (A) IN GENERAL. - The term 'specific learning disability' means a disorder in 1 or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in

imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations

(B) DISORDERS INCLUDED. - Such term includes such conditions as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia.

(C) DISORDERS NOT INCLUDED. - Such term does not include a learning problem that is primarily the result of visual, hearing, or motor disabilities, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage. 34 CFR 300.7(c) (10) (p.118)

Despite the multitude of definitions of MLD, there seems to be some consensus on how the disorder manifests. For instance, Gersten, Jordan, and Flojo (2005) explain that deficits in many number sense skills, such as counting and performing mental mathematics, are associated with MLD. More specifically, students with mathematics difficulties have poor counting strategies and struggle with retrieving basic math facts, hampering their ability to understand more complex algebraic concepts. The *DSM-IV-TR* also notes that a number of skills may be impaired in MLD including: linguistic skills involving naming mathematical codes or concepts or understanding written math problems, perceptual skills such as recognizing symbols and signs, attention skills such as paying attention to arithmetical signs, and mathematical skills such as following sequences, counting, and learning mathematics facts. Students with MLD may also show over-reliance on finger counting and demonstrate more errors when they are able to retrieve math facts (Hale & Fiorello, 2004).

MLD may be the result of a genetic predisposition, or various environmental or developmental factors (*DSM-IV-TR*; Geary, 2004). From a neuropsychological perspective, researchers have described several correlates of MLD (Hale & Fiorello, 2004; Mazzoco & Thompson, 2005). For instance, semantic memory may play a role in difficulty recalling facts and number symbol association. Visuo-spatial functions may be related to difficulties with place value and correctly lining up computation problems. Finally, frontal lobe functions such as executive functioning, attention, and working memory may be related to issues such as losing track of counting, reliance on finger counting, and monitoring problem solving steps.

In sum, a range of terminology exists to describe students with MLD. However, there is consistent indication that students with MLD will, at a minimum, display deficits in mathematics skills and mathematics achievement. Moreover, limited number sense knowledge and computation automaticity may differentiate students with MLD from typically performing students (Gersten et al., 2005). Further, as previously stated, students with MLD also progress more slowly than their non-disabled peers. Thus, failure to remediate difficulties associated with MLD has the potential to place students further at risk for a range of negative outcomes. In other words, the longer students with MLD go without intervention, the further they fall behind. Alternatively stated, the earlier students with MLD receive intervention, the greater their potential for more positive outcomes. Certainly, early intervention in mathematics appears to be a promising solution for students who experience MLD.

Early Academic Intervention in Mathematics

Policymakers, as reflected in recent legislation such as NCLB (2001) and IDEA (2004), as well as researchers (e.g., Gersten et al., 2005) have highlighted early intervention as one avenue for improving mathematics performance and increasing accountability, both on an individual student level as well as more generally. Several panels and consensus groups including the National Joint Council on Learning Disabilities (NJCLD, 2005), the National Council on Teaching Mathematics (NCTM, 2000), and the President's Commission on Special Education (2002) have emphasized prevention, and recommended screening, early identification, and early intervention with students with academic difficulties. Research supports these recommendations, demonstrating that early intervention can alter potentially negative outcomes (L.S. Fuchs & Fuchs, 2001). Progress monitoring, also described as formative assessment, is also included in the aforementioned recommendations as a way to determine whether students continue to benefit from core instruction, are responding to early intervention supports, and whether students need intensive individualized services or a reduction in these services. In sum, early intervention and formative evaluation of student progress represent two variables related to general increases in student achievement (Good & Brophy, 1986).

Research has begun to demonstrate that early intervention efforts targeting difficulties in mathematics have the potential to improve student learning (L.S. Fuchs & Fuchs, 2001). For example, Dev, Doyle, and Valente (2002) conducted a study with 11 six to seven year-old first grade students identified as at risk for reading and mathematics difficulties on the basis of low performance on individualized achievement tests. All

students participated in two commercially available intervention programs in mathematics and reading. The students showed significant gains in academics, so much so, that students considered at risk for reading and mathematics difficulties at the beginning of first grade no longer needed special education services by the end of the second grade.

Dowker (2005) also described two early intervention programs in mathematics, currently under development in England, for use with 6 to 7 year-old children. In the first program, *Mathematics Recovery* (Wright, Martland, & Stafford, 2000), teachers provide intensive individualized intervention, emphasizing counting and number representation, to low achieving students. The intervention occured daily for 30 minutes over a 12 to 14 week period. Teachers assessed students on key topics before and after the intervention. Although Dowker (2005) only provided a review of this program and did not indicate the specifics of research on the effectiveness program, preliminary evidence suggests significant student gains on topics that were problematic for students prior to the intervention. In the second program, *Numeracy Recovery* (Dowker 2001, 2003), teachers identified children having difficulty with arithmetic as targets for intervention. As reported by Dowker (2005), a pilot study of the Numeracy Recovery program is underway with 175 children who receive 30 minutes of weekly individual intervention for a period of approximately 30 weeks. Numeracy skills targeted for intervention are individualized for students based on their performance on pre-intervention assessment of nine areas of early numeracy such as counting, written symbolism of numbers, world problem solving, and arithmetical estimation. Again, preliminary evidence suggests that

students receiving this intervention, as compared to students in a control group, showed significant gains on post-intervention assessments.

Finally, in their review of early identification and intervention in mathematics, Gersten et al. (2005) also described several potential early intervention strategies that may benefit students with mathematics difficulties. For instance, Gersten et al. described that strategies from the *Number Worlds* curriculum (Griffin, 2004), such as having students practice counting upwards from a given number, counting backwards, and linking adding and subtracting to the manipulation of objects, can be easily implemented in classrooms and appear to help students build a sense of number. Gersten et al. also advocate for helping students develop a mathematics vocabulary.

Taken together, preliminary research on early intervention in mathematics illustrates that a number of intervention approaches and developing curricula may serve to improve students' mathematics performance. In particular, early intervention efforts seem to consistently emphasize counting principles, arithmetic skills, and mathematics vocabulary. However, it is also apparent that the study of early intervention in mathematics is in its early stages, with research often offering only preliminary findings. Still, it appears that when school personnel follow recommendations for early identification and intervention, they can generally expect positive outcomes for students.

Given that early intervention in mathematics demonstrates improved outcomes for students, it is important to identify students who would benefit from such programs. Furthermore, Clarke and Shinn (2004) have suggested that in order to maximize the effects of early intervention, school personnel should identify students at risk as soon as possible. In this way, it appears that the first step in making strides toward improving

student outcomes in mathematics is to identify students with early mathematics skills deficits. Not surprisingly, one common theme that pervades the research on early intervention in mathematics is the need for appropriate assessments to identify difficulties and monitor progress. In fact, Dowker (2005) points out that the most important conclusion that can be drawn from research on early intervention in mathematics is that it is critical to identify early indicators of mathematics difficulties in order to improve and prevent later mathematics difficulties. The theory of *number sense* has made contributions to understanding what such indicators may be.

Number Sense

Developmental psychology research has significant implications for determining which skills are critical to measure in order to identify students who might benefit from early intervention in mathematics. Early on, Piaget believed that several logical abilities were conditional to the development of arithmetic: seriation, classification, conservation of quantities, and inclusion (Piaget & Szeminska, 1941/1960). However, new research extends these principals. The theory of *number sense*, in particular, has become important in efforts to identify early mathematics difficulties. Investigators describe number sense in several different ways. Although no single definition exists, several commonalities exist among generally accepted definitions.

According to Gersten and Chard (1999), number sense refers to, "a child's fluidity and flexibility with numbers, the sense of what numbers mean, and an ability to perform mental mathematics and to look at the world and make comparisons" (p. 20). Gersten and Chard go on further to equate the relationship between number sense and mathematics to the relationship between phonics and reading. Kalchman, Moss, and Case (2001)

provided a second definition that states that number sense involves fluency in estimation, flexibility in mental computation, the ability to recognize unreasonable results, and the ability to transition among different representations and use the most appropriate representation.

Despite the lack of a mutually agreed upon definition of number sense, researchers have postulated several specific skills as representing the concept of number sense. These include, but are not limited to, counting, quantity discrimination, estimation, possessing a mental number line, and the ability to use multiple representations of the same number (Berch, 2005; Gersten et al., 2005). Gersten et al. add that factor analytic studies indicate that counting and quantity discrimination appear to be the two key factors involved in number sense and that these two skills serve as precursors to other number sense skills such as estimation and using multiple representations. Gersten, Clarke, and Jordan (2007) also emphasize that the ability to make quantity comparisons is central to number sense.

Building on theoretical and operational definitions, the theory assumes that children's number sense is the foundation building block of all other math knowledge. For example, Shapiro (2004) describes number sense as a necessary prerequisite skill to facilitate basic mathematics computational knowledge. Also, in a paper outlining their conceptualization of number sense, Gersten and Chard (1999) assert that number sense leads to automaticity in mathematics and is crucial to students' ability to solve basic arithmetic computations. Gersten et al. (2005) echo this notion and state that conceptual linkages associated with number sense are necessary tools for assisting students with thinking about mathematics problems and developing higher order thinking for working

on mathematical problems. Interestingly, many of the skills associated with number sense are related to the hallmarks of mathematics difficulties as described in the previous section on MLD. Again, students with MLD show deficits in counting skills and strategies, perceptual skills related to recognizing symbols and signs, sequencing, and fact retrieval (Gersten et al., 2005; APA, 2004). This connection makes the case for the assessment of number sense skills in early intervention efforts even more compelling.

Thus, while a growing body of research underscores the importance of examining number sense skills as part of early intervention efforts; these assessment methods vary. The following section more fully explains assessment approaches and tools used to identify students with mathematics difficulties. In particular, the section emphasizes CBM used in a formative evaluation framework as a potentially valuable approach to the early identification of mathematics skills deficits.

Assessment Approaches

Assessment is a core component of educational practice and some researchers define it as the process of collecting data in order to understand students' problems and make individualized educational decisions (Salvia and Ysseldyke, 2001). Salvia and Ysseldyke identify five types of decisions that one can make from assessment: (a) referral, (b) screening, (c) classification, (d) instructional planning, and (e) monitoring students' progress. In addition, they note that school personnel can use assessment data to determine the effectiveness of educational programs or interventions. Shapiro (2004) explains that it is critical for decision making to match assessment data. For example, Shapiro explains that norm-referenced instruments may be useful for classification, while criterion-referenced measures are better matched to decisions about relative strengths and

weaknesses in academic skills. CBM has particular relevance to progress monitoring and decision making about students' responsiveness to interventions.

Psychological and educational assessment has traditionally relied on the use of norm-referenced, standardized cognitive and achievement tests to assess and diagnose LD. More specifically, D. Fuchs, Mock, Morgan, and Young (2003) explain that educators came to characterize LD as a severe discrepancy between performance on cognitive (IQ) and achievement tests, which is consistent with current diagnostic criteria set forth in the *DSM-IV-TR*. As such, most state departments of education adopted the severe discrepancy principle, or the IQ-Achievement discrepancy approach, as the basis for determining LD and the subsequent provision of special education services. The primary purpose of the discrepancy approach, then, has been to identify students who meet criteria for LD. This approach has been successful in that it has allowed many students to receive services that support their learning needs. Outside of their use for determining discrepancies between cognitive functioning and academic performance, commonly used norm-referenced assessment tools are beneficial in that they provide information on students' relative performance compared to large samples of their same-aged peers and assess a wide range of skills.

However, investigators and school personnel have advocated for alternative approaches to using the IQ-Achievement discrepancy approach for identifying academic difficulties. For instance, one criticism is that this approach is highly inconsistent across states (D. Fuchs et al., 2003; Fletcher et al. 2004; NJCLD, 2005). First, there are inconsistencies with respect to how one computes discrepancy, for example by either subtracting a student's standard achievement score from his or her standard IQ score, or

by examining the regression of IQ on achievement. Secondly, the size of discrepancy required to make a determination of LD is inconsistent, ranging from 1 to 2 standard deviations. Finally, the specific tests used to measure intelligence and academic achievement also vary.

In addition, the NJCLD (2005) and others (e.g., D. Fuchs et al. 2003; Fletcher et al. 2004) have argued that the discrepancy approach is essentially a "wait-to-fail" model where students must demonstrate poor performance for years before their achievement scores fall significantly below their IQ scores. In other words, this identification approach may not be providing necessary support services to students until they are performing substantially behind their peers. Therefore, helping these children catch up to their classmates is made more challenging and often results in special education serving as an end point rather than a gateway to more individualized appropriate instruction (e.g., Fletcher et al., 2004; NJCLD, 2005). Finally, it also stands to reason that this approach excludes low achieving students who may not demonstrate significant discrepancies, but are no less deserving of academic support services. This is especially true as research (e.g., Fletcher et al., 1994) suggests that the performance of low achieving students and students diagnosed with an LD on the basis of an IQ-Achievement discrepancy cannot be reliably differentiated, at least in the area of reading.

In the current educational context, where there is a premium placed on early identification and intervention, the discrepancy approach for identifying LD is losing standing to other procedures that make support services available for children in a more timely fashion. Changes in educational legislation and research reflect the criticisms of more traditional assessment approaches (e.g., D. Fuchs et al., 2003). Response to

Intervention (RTI) has become a viable service provision framework for students with disabilities, as the most recent reauthorization IDEA (2004) indicated that RTI may be used to assess and monitor students with learning difficulties in addition to, or in place of, former practices involving examining discrepancies between students' cognitive abilities and academic performance, as measured by standardized, norm-referenced measures. In a similar vein, the report from the NJCLD (2005) recommends expanding the notion of assessment to include formative methods that screen students for academic difficulties and monitor performance across all children, as is done when using a multi-tiered model of service delivery such as RTI. D. Fuchs et al. also note that many professional organizations support RTI approaches including the Division for Learning Disabilities of the Council for Exceptional Children, the International Dyslexia Association, the National Association of School Psychologists, and the National Association of State Directors of Special Education, to name a few.

The most recent reauthorization of IDEA (2004) introduced RTI in an attempt to reduce many of the problems associated with the IQ-Achievement discrepancy model. Namely, RTI has the potential to help more students in a timelier manner, it provides more individualized and intensive instruction to students, it has the potential to reduce special education enrollment and associated costs, and it provides services that do not depend on IQ test performance (D. Fuchs et al.). In this way, RTI presents several advantages over the IQ-Achievement discrepancy model.

RTI is a multi-tiered framework of service delivery. Consistent with the requirements of NCLB (2001), RTI involves ensuring the use of empirically supported curricula and intervention programs as well as screening and monitoring student progress.

Several sources outline the procedures required by RTI (e.g., D. Fuchs et al., 2003; D. Fuchs, 2001; NASP, 2007; NJCLD, 2005). Specifically, RTI requires that, at the first tier of service delivery, classroom teachers provide all students with universal high quality instruction and undergo ongoing progress monitoring. In the second tier, teachers provide students who do not make adequate progress, or do not respond in the first tier, with instruction targeting skill weaknesses in small groups. Finally, in the third tier, students who still fail to make progress should receive individualized interventions that may consist of special education. At this stage, the school psychologist is likely to implement a comprehensive special education evaluation to provide a broader view of student difficulties.

Types of RTI models. Although the general principles remain the same, school personnel may implement RTI in various ways, with different numbers of levels, or tiers, in the process or with different service providers. One specific way to implement RTI that is most frequently used in schools and supported by behaviorally-oriented school psychologists is called the Problem-Solving Model (D. Fuchs et al., 2003). In this model, which is based on the principles of behavioral consultation, the school psychologist uses an inductive, empirical, and behavioral approach to identify difficulties and monitor students' performance in school. Although the Problem-Solving Model may differ among districts or schools, necessary stages of the model include problem identification, problem analysis, plan implementation, and problem evaluation (D. Fuchs et al.). Quantitative data, usually consisting of CBM, measuring students' initial performance (baseline) and subsequent responses to empirically supported interventions, support each

stage. In this way, the Problem-Solving Model utilizes data for educational decision making.

In addition to the Problem-Solving Model, D. Fuchs and Fuchs (2001) provide a description of another set of procedures that constitute another way to implement RTI. This model consists of three tiers of intervention and involves four stages. In the first stage, school psychologists assess all students in a school using brief screening tools that demonstrate the ability to predict performance on high-stakes state assessments or local graduation requirements. Alternatively, the first stage could involve individual screening of students who performed below the 25th percentile on the previous year's state assessment or on current achievement tests. The second stage is two-fold and involves exposing all students to evidence-based curricula and instruction (tier 1 intervention) and continued progress monitoring using brief assessments. D. Fuchs and Fuchs expect approximately 80% of students to be successful at this level of intervention. In the third stage, teachers provide students nonresponsive students to tier 1 interventions with small group instruction (tier 2 intervention) and continuous progress monitoring with brief assessment tools. It is expected that 10-15% of students will require this level of intervention. Finally, in the fourth stage, students who failed to respond in the third stage receive a comprehensive special education evaluation (tier 3) under the procedural safeguards outlined in IDEA (2004). Fuchs and Fuchs expect approximately 1-5% of the school aged population is expected to require tier 3 level supports. This model is illustrated in Figure 1.

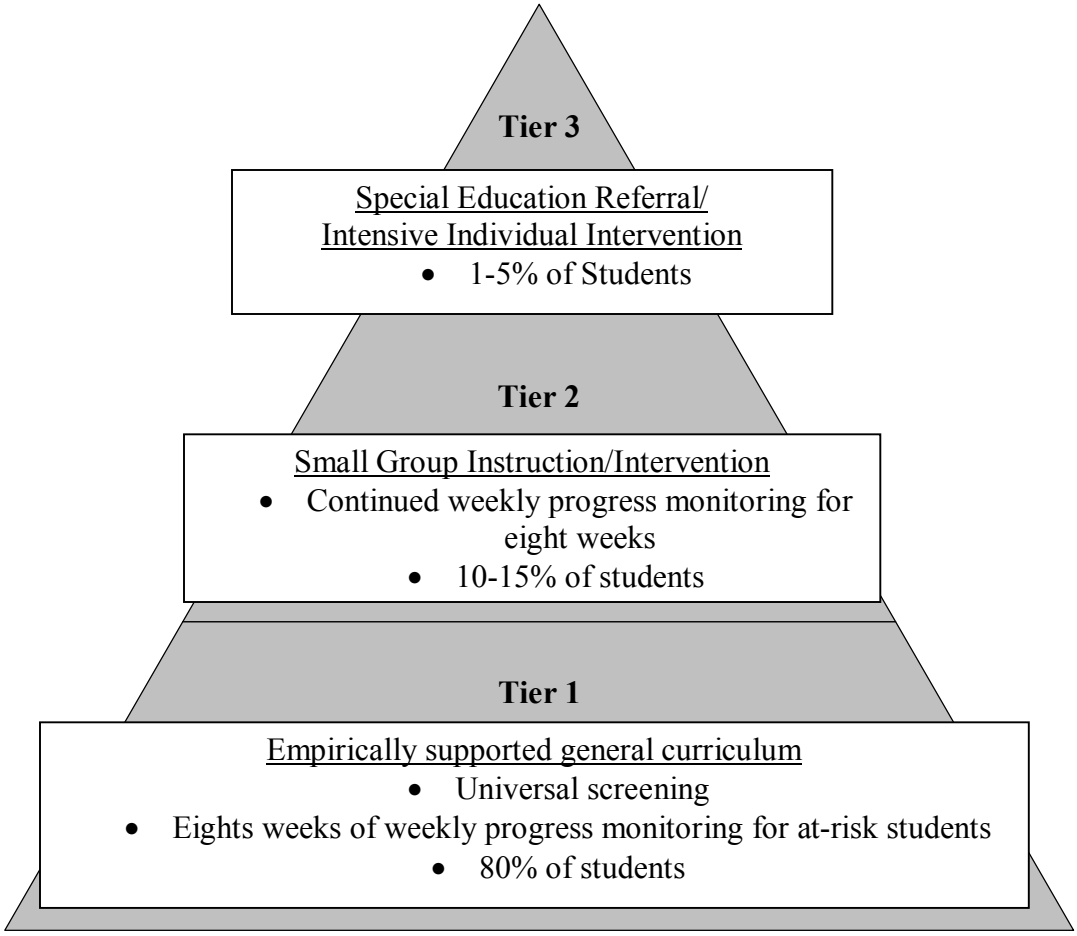


Figure 1. Response to Intervention Model based on D. Fuchs and Fuchs (2001)

Screening and progress-monitoring decisions. Due to the necessity for screening and frequent data collection on students' progress in the curriculum, RTI requires the use of brief, standardized measures that are matched to the curriculum. Yet, as indicated earlier, many standardized, norm-referenced assessments do not fit this description. Specifically, several researchers (e.g., Bieber & Choi, 2004; Deno, 1992; Shapiro, 2004) point out that norm-referenced achievement tests are not always related to the curriculum

used in particular schools, do not directly relate to educational interventions, and are not a necessarily efficient way to measure short-term, individual student progress. Also, Gersten et al. (2005) state that achievement tests have many different types of items and therefore, may mask specific deficits. Therefore, while proponents of alternative assessment approaches will acknowledge that many standardized, norm-referenced achievement and cognitive tests are well designed and provide useful data; they argue that they have limited utility for progress monitoring and intervention planning (Shinn & Bamonto, 1998). In contrast, CBM is specifically designed for such purposes.

Curriculum-Based Measurement

When one combines the criticisms of the discrepancy approach, the call for simplified assessment techniques matched to curriculum, and the increasing support for an RTI model, the need for tools matched to decisions about planning interventions and monitoring progress is evident. Curriculum-Based Assessment (CBA) is one type of assessment that may be used to facilitate the early identification and monitoring of students with academic difficulties in an RTI framework. CBA refers to the direct standardized assessment of basic academic skills (Shapiro, 2004). Different models of CBA exist, including criterion-based and accuracy-based models (Shinn & Bamonto, 1998). While these approaches are somewhat different, they share similar theoretical and practical characteristics. The key common characteristics of the various models of CBA are that assessment practices are tied to instructional interventions, are brief, and may be used to monitor student progress and the effects of instruction (Bieber & Choi, 2004; Shapiro, 2004).

Whereas CBA is a general term, CBM refers to a specific, research-supported approach to student progress monitoring that Stanley Deno and researchers at the University of Minnesota Institute for Research on Learning Disabilities developed between 1977 and 1983 (Deno, 1992). Originally, the intent of Deno's work was to provide special education teachers with a tool to accurately and efficiently assess the effects of their instruction (Shinn & Bamonto, 1998). The research literature, spanning over 40 years, supports CBM as a reliable and valid assessment approach for evaluating elementary level academic skills (Shapiro, 2004; Shinn, 1989). In fact, results of hundreds of research studies of CBM support this tool, which researchers once denoted as an emerging alternative, as a validated alternative to traditional assessments (L.S. Fuchs, 2004; Shapiro, 2004).

Shinn and Bamonto (1998) provide an overview of the specific procedures used in CBM. First, the authors explain that CBM refers to a set of "standard simple, short-duration fluency measures of reading, spelling, written expression, and mathematics computation" (p. 1). That is, CBM measures important basic skills relevant to school achievement. The primary testing strategies involved in CBM vary depending on the subject area. In general, school personnel conduct CBM via frequently administered grade-level skill measures, called probes, which typically range from 1 to 5 minutes in length. In mathematics, CBM involves the administration of 2 to 5 minute probes where students write answers to computational problems and the number of digits correct per minute serves as the measure of performance.

When considered overall, the results of CBM can serve as vital signs of student achievement in basic academic skills (Shinn & Bamonto, 1998). Shinn and Bamonto

further describe the "big ideas" necessary to understanding how to use the results of student performance on CBM. The central "big idea" that these authors discuss is the fact that CBMs are validated for use as "dynamic indicators of basic skills" (p. 5). The dynamic nature of CBM allows for measuring differences both among individuals and within individuals over time. The notion that CBM is a skills indicator implies the validation of CBM as a correlate of behaviors indicative of overall performance in various academic areas. In other words, researchers have validated CBM data consisting of the number of words read correctly in one minute as an indicator of overall reading achievement; for example, through research demonstrating that CBM performance successfully predicts outcomes on various broad academic criterion measures (Deno, 1992). Finally, CBM is also a valid measure of specific basic skills in reading, mathematics computation, written expression, and spelling as evidenced by concurrent, criterion- related validity correlations with other standardized academic measures (e.g., Marston, 1989).

Other big ideas include: (a) the principal purpose of CBM is formative evaluation, and (b) CBM is central to the RTI framework for making educational decisions about individuals (Shinn & Bamonto, 1998). Specifically, school personnel can use CBM to identify a problem, clarify the problem, and measure the students' progress after the implementation of an intervention (e.g., D. Fuchs & Fuchs, 2001). CBM facilitates formative evaluation and frequent monitoring of student process because it uses standardized procedures, is logistically feasible, and is sensitive to improvement over time (Shinn & Bamonto, 1998). The use of standardized procedures ensures that one can attribute changes in performance to actual growth rather than changes in assessment

procedures. The brief and efficient nature of CBM makes the frequent assessment of performance required in formative evaluation feasible. Finally, the sensitivity of CBM to reliably detect changes in student growth, allows school personnel to make decisions about student progress.

In summary, CBM represents a procedure that one can readily use in an RTI framework and offers several assessment advantages that are worth noting. First, because it is simple to administer, generally, most school personnel, not just those with specialist degrees, can effect CBM. Second, CBM addresses many of the criticisms of traditional assessment procedures. Namely, CBM is linked to the curriculum and instructional interventions, sensitive to small changes over time, and uses consistent procedures. These properties permit frequent analysis of students' skill mastery and progress toward short-term and year-end curricular objectives. Third, the brevity of CBM allows for rapid documentation of student progress and relatively quick decision making. Fourth, more recent research has shown that CBM may also be used to predict student outcomes on high-stakes testing and to measure growth in secondary and early childhood programs. For example, Shapiro, Keller, Lutz, Santoro, and Hintze (2006) examined the relationship of CBM of reading, math computation, and math concepts and applications with either performance on a statewide standardized achievement test, the Stanford Achievement Test-Ninth Edition (Harcourt Brace Educational Measurement, 1996), Metropolitan Achievement Test-Eighth Edition (Harcourt Brace Educational Measurement, 2002), or performance on the Stanford Diagnostic Reading Test (Karlsen & Gardner, 1995). Approximately 1400 students in two rural districts in Pennsylvania participated in the study. Shapiro et al. found that both reading and math CBM had moderate to strong

correlations with both types of standardized tests and concluded that CBM could serve as an indicator of future performance on standardized achievement tests.

While the advantages of CBM are apparent, CBM also has limitations. Specifically, it is critical to understand that CBM is solely proposed to be used as a measure of basic skills, typically at the elementary school level. That is, CBM is not presently designed to measure knowledge in specialized content areas such as history, literature, or other academic areas that students may be exposed to past the elementary level. Similarly, although CBM is well-matched to the RTI model; it is not intended to replace other measures that may be used to ultimately make decisions about the presence of various disabilities and students' eligibility for special education services. Rather, it is recommended as a critical part of a multiple-assessment process that is recommended within the RTI model. To summarize in the context of Salvia and Ysseldyke's (2001) assessment model, research supports CBM as a tool to assist with academic decisions related to screening, instructional planning, and progress monitoring.

History of M-CBM Research. The research literature has supported CBM for some time and the study of CBM has gone through various stages. Following initial development efforts, researchers examined CBM in terms of its technical properties and logistical features (L.S. Fuchs, 2004; Shinn & Bamoto, 1998). For example, researchers have investigated the technical features of single-scores obtained from CBM at one point in time (static scores), the technical features of scores obtained from CBM over time (slope), and more recently, the instructional utility of CBM (L.S. Fuchs). As noted by Marston (1989) and indicated in ethical guidelines set forth by professional associations such as the American Psychological Association and the American Educational Research

Association, such research was essential because assessment procedures need to demonstrate evidence of reliability and validity. L.S. Fuchs and Good and Jefferson (1998) indicate that the data available on validity of CBM is vast, so much so, that even summaries of the validity of CBM contain overwhelming amounts of information. Thus, while a comprehensive review of the technical features of CBM is beyond the scope of this paper, summaries of key findings on mathematics-CBM (M-CBM) follow.

First, in a review of the research on progress monitoring in mathematics, Foegen et al. (2007) note that M-CBM demonstrates high estimates of test-retest, parallel forms, and inter-rater reliabilities, often with correlations of .80 or above. In one specific study, Tindal, Marston, and Deno (1983) examined the reliability of CBM probes based on data from a sample of 566 randomly selected students in grades 1 through 6. Analysis of the inter-rater reliability of several math computation CBM probes (e.g. addition, subtraction) yielded coefficients ranging from .90 to .99. In another study, Thurber, Shinn and Smolkowski (2002) determined that interscorer reliability was .83.

With respect to validity, CBM of math skills does not always demonstrate as strong a relationship to standardized math assessments, with estimates in the .50 to .70 range (Foegen et al., 2007). This may be partly attributed to concerns that available math tests serve as poor criterion measures and measure skills beyond mathematics computation (e.g., reading skills) (Marston, 1989). However, Shinn and Marston (1985) demonstrated construct validity for M-CBM. These authors found that, in a study comparing the academic performance of general education students and those receiving different levels of special education services, performance on basic multiplication and division probes reliably differentiated students in different programs. Another study

examined the relationship between M-CBM (i.e., computation) probes with math applications measures that assessed students' ability to apply basic math facts and principles. They found that computations and applications were separate, but highly related, skills and that M-CBM is a measure of computation (Thurber et al., 2002). Finally, with respect to slope research, or research directed at understanding student progress, studies show that CBM is able to detect gains in student performance (Shinn & Bamonto, 1998; Shapiro, Edwards, & Zigmond, 2005).

Following research on the technical properties and sensitivity of CBM, the role of CBM in educational decision making, or in terms of instructional utility, became the central area of study. Foegen et al. (2007) report that existing studies support the use of progress monitoring measures to improve student achievement. Marston (1989) also describes research findings on the role of CBM in this respect. First, Marston notes that CBM performance reliably differentiates students placed in special education programs from those in regular education. In the study described previously, Shinn and Marston (1985) provided specific data supporting this assertion. In another study, Stecker and Fuchs (2000) found that when teachers made instructional adjustments based on CBM progress monitoring data, students performed significantly better on a global achievement test than did peers whose adjustments were not based on such data.

Early Numeracy CBM

In addition to its utility for assessing elementary level academic skills, several authors have recently proposed CBM as a tool for use with early identification and intervention of beginning academic skills (Daly et al., 1997; VanDerHeyden et al., 2001; VanDerHeyden et al., 2004). Until very recently, research on the use of CBM with early

grade (i.e., kindergarten and first grade) students has focused on reading. Gersten et al. (2005) indicate that over 20 years of theory building and research have lead to the development of reliable and valid screening measures of beginning reading skills; while the development of similar mathematics measures "is in its infancy." (p. 293). Shinn (1989; 1998) also reports that, although there is support for the use of reading CBM in the early grades, less research exists to support the use of M-CBM in the early grades.

Investigation of M-CBM at the early grade levels is important as it may help schools to address accountability requirements, which are explicitly mandated in legislation such as NCLB (2001), as well as assist schools with early intervention efforts. Because, as described earlier, CBM facilitates evidence based practice (Shinn, 1989), is brief and relatively simple to administer, and is a method of formative evaluation, it might also be used to study response to intervention with early grade students. These benefits of CBM correspond with recommendations that identification of academic difficulties should be simplified and student progress be monitored over time with the intent to make data-based decisions (President's Commission on Special Education, 2002; IDEA, 2004). Even more specifically, the use of M-CBM for assessing early academic skills is in line with recommendations for the early identification of students in need of academic intervention and would potentially meet the critical need for early indicators of mathematics difficulties (Dowker, 2005).

As researchers began studying the use of M-CBM for early identification and early intervention efforts, one particular problem with using M-CBM for assessing early math skills arose. As computational problems compose M-CBM probes, they require that teachers have exposed students to mathematics computation instruction. In this way, M-

CBM would not be useful for students below the first grade level because they have not had sufficient mathematics computation instruction. Not surprisingly, Clark and Shinn (2004) note that students below the first grade level will often attain scores of zero on M-CBM. Consequently, CBM probes based on prerequisite mathematics skills, consistent with educational objectives for younger children, and based on research associated with early number knowledge have emerged (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Daly et al., 1997; VanDerHeyden et al., 2001; VanDerHeyden et al., 2004; VanDerHeyden et al., 2006).

There are several rationales for selection of number sense skills as the content of early indicators of mathematics abilities. Gersten and Chard (1999) explain that students often informally acquire number sense before formal schooling begins. In this way, number sense skills appear to be a natural choice of content in math assessments for young children. Also, because number sense is necessary for learning formal arithmetic in the early elementary grades and also relates empirically to learning disabilities; one would expect measures based on number sense to predict later mathematics difficulties (Gersten & Chard, 1999; Griffin at al., 1994). In other words, one would expect measures of number sense to demonstrate validity for use as screening measures for mathematics difficulties. In fact, in a study with 200 kindergarten children in two urban areas, Baker, Gersten, Flojo et al. (2002) found that performance on the Number Knowledge Test (Okamoto & Case, 1996) correlated with subsequent performance on mathematics subtests of standardized achievement measures. In another study, Jordan, Kaplan, Locuniak, and Ramineni (2007) found that a number sense battery, consisting of measures of skills such as counting, number knowledge, nonverbal calculation, number

combinations, and word problems, administered in the beginning of kindergarten correlated strongly with math achievement, as indicated by the Calculation and Applied Problems subtests of the Woodcock-Johnson III (McGrew &Woodcock, 2001), in the middle of first grade.

Researchers (e.g., Clarke & Shinn, 2004) have used this theoretical basis to develop CBMs that serve as indicators of students' number sense. Stated otherwise, the content of such measures draws on number sense theory with format and procedures belonging to CBM. Research supports this line of development noting that better operationalized definitions of number sense will provide the bases for continued advances in developing valid early screening and detection measures for mathematics difficulties (Gersten et al., 2005; VanDerHeyden et al., 2004).

Current research findings. Preliminary work in the area of CBM of early math skills has examined measures involving several different types of mathematics tasks including: number reading, counting, writing numbers, selecting numbers, and drawing a number of objects given a specific number (Daly et al., 1997; VanDerHeyden et al., 2001) with children in preschool through first grade.

Beginning with studies with preschool children, VanDerHeyden et al. (2004) examined a set of six CBMs of early mathematics with 102 preschool students in the Southern U.S. Students were from seven classrooms in two rural preschool programs. In one of the preschool programs, 55% of the students were African-American, 43% were Caucasian, and 2% were described as "Other." Fifty-seven percent of these students were male. In the second program, 56% of the students were Caucasian, 36% were African-American, 5% were Hispanic/Latino, and 3% identified as "Other." Sixty-four percent of

the students in the second program were male. The measures examined skills including counting objects, selecting numbers, naming numbers, selecting shapes, counting, and visual discrimination; and state and local performance standards served as the basis of their development. On the counting objects measure, researchers timed students as they counted pictures presented to them on laminated cards. The selecting numbers task required students to point to a specific number stated by the examiner when provided with a laminated card containing four different numbers on it. The naming number task required students to name a single number presented to them on a laminated card. Similar to the selecting number task, the selecting shapes task presented four shapes on a laminated card to students and required them to point to the shape named by the examiner. The counting task simply required children to count aloud while being timed. Finally, the visual discrimination task presented students with a laminated card with four objects (e.g., numbers, letters, shapes) and required students to point to the one object that differed from the other three. In this study, researchers administered the measures individually.

VanDerHeyden et al. (2004) found that four of the measures (choose number, number naming, count objects, and object discrimination) demonstrated adequate alternate forms reliability (r = .83 to r = .88). These four measures also demonstrated moderate correlations with the Test of Early Mathematics Ability (TEMA-2; Ginsburg & Baroody, 1990), a measure of math performance (r = .39 to r = .52) and the Brigance Screens (Brigance, 1985), a standardized screening measure of early academic skills (r = .44 to r = .57)

Later, VanDerHeyden et al. (2006) examined the progress monitoring and screening accuracy of the four preschool measures that performed well in her previous research (choose number, number naming, count objects, and object discrimination) from preschool to kindergarten. Participants consisted of the same preschool children from VanDerHeyden et al.'s 2004 study. In addition, the authors administered three kindergarten measures. The kindergarten probes measured visual discrimination, counting objects, and circling letter skills. On the kindergarten discrimination measure, VanDerHeyden et al. showed students four items (e.g., letters, shapes, numbers) that matched and one that did not match and required them to select the object that did not match. The kindergarten circle number measure required students to count a set of circles on one side of a page and circle the correct corresponding number from a list of four number choices on the other side of the page. Finally, the kindergarten circle letter measure required students to circle letters corresponding to the beginning letter sound given a picture of an object and having the examiner state the name of the object. Overall, performance on preschool measures correlated moderately with performance on kindergarten measures with correlations ranging from .40 to .60. VanDerHeyden et al. also obtained preliminary evidence that preschool measures accurately identified children needing intervention by comparing CBM performance to scores on the Brigance Screens (Brigance, 1985), a standardized, norm-referenced screening device for children aged 2 to 5 years-old.

More recently, Floyd, Hojonski, and Key (2006) developed and evaluated another set of early numeracy measures for preschool children. Floyd et al. described preliminary reliability and validity data for a set of measures called the Preschool Numeracy

Indicators (PNIs). The authors individually administered the PNIs and three criterion measures to a sample of 163 3-to-6 year old children attending four preschool education settings. As a result of the authors' review of texts on number skills development, they selected standards for early childhood mathematics education and Head Start and the principles of number and operations as the focus of the PNIs. Specifically, the PNIs measured skills such as counting objects, oral counting fluency, number naming fluency, and quantity comparison fluency. When Floyd et al. examined reliability and validity of the PNIs, they found evidence for moderate to strong internal relationships between the four PNIs, suggesting that each measured a single underlying construct. The PNIs also demonstrated acceptable test-retest reliability estimates for all but the measure involving counting objects, and moderate relationships with existing validated measures of school readiness and early number and mathematics skills.

Another set of research has examined first grade measures. Daly et al. (1997) examined the interscorer and test-retest reliability and the criterion-related concurrent and predictive validity of 11 CBMs that targeted skills central to the first grade curriculum with a sample of 30 first grade students in an urban school setting. The participants consisted of 25 African-American, 4 Caucasian, and 1 Latino students. Sixteen of the children were boys and 14 were girls. Math skills assessed by Daly et al. included number reading and writing, counting, and selecting numbers. More specifically, the number reading measure required students to read randomly arranged numbers up to 20. The number writing measure required students to write numbers up to 20 given dictated numbers from the examiner. On the number counting measure, students simply counted aloud for 1 minute. Finally, on the number selection measure, the examiner dictated

numbers to students and required the students to circle the corresponding number on a sheet of paper with randomly arranged numbers. The examiner individually administered number reading and number counting measures, while he or she administered the number writing and number selection measures in groups of 15 students. Daly et al. used Pearson correlations to analyze the test-retest reliability and criterion and predictive validity of the measures in their study.

First, with respect to reliability, Daly et al. (1997) found that all the mathematics measures demonstrated adequate interscorer reliability, as determined by the number of agreed upon responses divided by the number of agreed upon and disagreed upon responses, with percentages ranging from 86% to 100%. Yet, only the number reading and number counting measures demonstrated adequate test-retest reliability (r = .82, and r = .88, respectively). Daly et al. determined concurrent criterion-related validity by examining relationships among the experimental measures and broad reading and math scores on the Woodcock Johnson Test of Achievement-Revised. In this way, Daly et al. only found moderate correlations between the number counting measure and broad math scores on the Woodcock Johnson-Revised. Finally, they determined predictive validity by examining relationships among the experimental CBM measures and later performance on mathematics and reading CBM. Interestingly, while the number reading and number counting measures demonstrated moderate to high correlations (r = .44 - .49) with reading CBM; unfortunately, none of the mathematics CBMs demonstrated significant predictive validity for M-CBM.

Clarke and Shinn (2004) developed another set of early mathematics indicators that are commercially available on AIMSweb®, a web-based progress monitoring and

RTI system that provides CBM assessment materials. Clarke and Shinn used preliminary data to assess reliability, validity, and sensitivity with four measures designed for kindergarten and first grade students. The authors developed these individually administered CBMs based on the principles of number sense and refer to these measures collectively as the Tests of Early Numeracy (TEN). The TEN measures correspond with number sense skills such as counting, identifying numbers, quantity discrimination, and possessing a mental number line (i.e., missing number). These tasks require students to count aloud, name numbers presented to them on a sheet of paper, choose the larger number of two numbers presented in boxes on a one-page grid, and name the number missing from a string of 3 numbers presented in boxes on a one-page grid.

Clarke and Shinn (2004) administered their four TEN measures to 52 first grade students from the Pacific Northwest. The majority of students in the study were Caucasian with 2 students who were Native American and 3 who were Hispanic. Twenty-nine of the participants were female and 23 were male. Clarke and Shinn examined the test-retest, alternate form, and interscorer reliability of TEN as well as the concurrent and predictive validity of the measures with a sample of first grade students by administering assessments in the fall, winter, and spring of one school year. In general, the authors found high coefficients for all of the measures for each type of reliability and validity evidence.

Specifically, Clarke and Shinn (2004) reported that inter-scorer reliability ranged from .98 to .99, whereas alternate form reliability ranged from .78 to .99, and test-retest reliability ranged from .79 to .86. Concurrent validity correlations among the experimental measures and with M-CBM and a standardized measure of mathematics

problem solving *(*The Woodcock Johnson-III Applied Problems subtest; Woodcock, McGrew, & Mather, 2001) ranged from .50 to .93. Predictive validity correlations ranged from .46 to .78 with respect to math computation as measured by first grade M-CBM and from .68 to .79 with respect to students' performance on The Woodcock Johnson-III Applied Problems subtest. Furthermore, TEN measures were sensitive to student growth over time. According to the results, quantity discrimination showed the most support for use as a single indicator of early mathematics, while the least support was for counting, although there was still support for counting as a sole indicator of early math skills. In this way, the quantity discrimination measure might provide educators with the most effective screening device for determining which first grade students may develop difficulties with arithmetic and problem solving.

Finally, research has examined kindergarten measures. Chard et al. (2005) extended the work of Clarke and Shinn (2004) in their research on CBM of number sense skills. Specifically, Chard et al. extended previous research on TEN by investigating the concurrent and predictive validity of TEN with a sample of kindergarten and first-grade students. Chard et al. used three of the same experimental TEN measures as those used by Clarke and Shinn (identifying numbers, quantity discrimination, and identifying missing numbers), with some modifications. Specifically, Chard et al. modified the measures to only include numbers up to 10 at the kindergarten level because kindergarten students may not have been exposed to numbers 10 to 20. In addition, Chard et al. included additional number sense measures and one criterion measure of number sense-the Number Knowledge Test (Okamoto & Case, 1996). They administered the measures individually to participants from the fall to the spring of one academic year. The

experimental number sense measures used by Chard et al. included the following tasks: (a) counting aloud to 20, (b) counting up aloud from 3 and 6, (c) counting aloud by 2, 5, and 10, (d) orally identifying numbers from 1 to 20 at the first grade level and 1 to 10 at the kindergarten level, and (d) writing a number corresponding with one orally provided by an examiner.

Findings from Chard et al.'s (2005) were consistent with Clarke and Shinn's (2004) research by replicating concurrent and predictive validity of TEN with first graders and providing preliminary evidence of TEN's effectiveness with kindergarten students. With the first grade students, validity coefficients ranged from .45 to .61. These findings approximated those of Clarke and Shinn, although Chard et al.'s correlations were somewhat smaller. With kindergarten students, validity coefficients for the experimental TEN measures ranged from .50 to .63. Furthermore, each of the TEN measures demonstrated sensitivity over the school year.

VanDerHeyden et al. (2001) also examined the reliability and validity of group administered CBM readiness probes in reading, mathematics, and writing for kindergarten students. They designed these measures to identify kindergarten students with deficient readiness skills. The math measures examined by VanDerHeyden et al. (2001) assessed counting skills, circling numbers, writing numbers, and drawing objects given a specific number. In particular, the circle number measure required students to count a set of circles on one side of a page and then circle the correct number from a list of possible choices on the other side of the page. The number of circles ranged from 1 to 10 and the measure provided students with four number choices. The write number measure required students to count a set of objects and write the corresponding number in

a box. Again, the number of objects ranged from 1 to 10. The drawing objects task provided students a series of numbers ranging from 1 to 10 and required them to then draw a corresponding number of circles. All measures were group administered.

VanDerHeyden et al. (2001) administered the measures to 107 kindergarten students from suburban public schools in the Southern U.S. More specifically, students from two schools comprised the participant sample. One of the schools had a student population that was 70% Caucasian, 29% African-American, and 1% described as "Other", while the other school's population consisted of 72% African-American, 25% Caucasian, and 3% "Other" students. Fifty students were female and 57 were male. The authors found adequate reliability for screening purposes for three of the six measures examined. Specifically, with respect to the internal consistency of the math measures, the circle number and write number measures yielded coefficient alphas that surpassed .90. The draw circles measure yielded a coefficient alpha just above .80. In addition, alternate form correlations for the circle number and write number measures exceeded .80. With respect to the validity of the measures, VanDerHeyden et al. found moderate correlations for the three math readiness measures and math composite scores on a criterion measure (the Comprehensive Inventory of Basic Skills, Revised [CIBS-R], Brigance, 1999). Scores also predicted which students were retained in kindergarten or promoted to first grade. VanDerHeyden et al. (2001) concluded that CBM readiness measures had the potential to serve as a useful adjunct to other individually administered assessments, although there still needs to be evidence for the sensitivity of these measures to monitor progress over time.

Together, studies exploring early numeracy CBM provide evidence for a range of

measures that may serve as useful screening and progress monitoring tools for preschool through first grade students. At the preschool level, measures involving choosing numbers, number naming, counting objects, object discrimination, oral counting fluency, number naming fluency, and quantity comparison fluency have demonstrated adequate technical properties (VanDerHeyden et al., 2004; Floyd et al., 2006). In addition, there is also evidence that preschool measures of choosing numbers, naming numbers, counting objects, and object discrimination can accurately predict which preschool students will need intervention in kindergarten (VanDerHeyden et al., 2006). First grade measures of number reading and counting demonstrate adequate test-retest reliability, with the number counting measure also demonstrating moderate concurrent validity relationships with other measures (Daly et al., 1997). Clarke and Shinn (2004), along with Chard et al. (2005), also found adequate technical properties for first grade level TEN, which measures oral counting, number identification, quantity discrimination, and number line skills. Finally, there is also adequate technical support for measures used at the kindergarten level. These include kindergarten level TEN measures (Chard et al., 2005) and VanDerHeyden et al.'s (2001) measures of counting skills, circling numbers, writing numbers, and drawing objects given a specific number.

This dissertation focused on the TEN for several reasons. First, there is more research available verifying the technical properties of the TEN than there is for other measures. Studies examining TEN (e.g., Clarke & Shinn, 2004; Chard et al., 2005) indicate that these measures are reliable, valid, and sensitive to growth for first grade students. Research also demonstrates validity and sensitivity to growth for kindergarten students (Chard et al., 2005). This research demonstrated support for a measure of

students' ability to discriminate quantities and identify the missing number in a string of numbers for both kindergarten and first grade students. In addition, counting skills also seem to be an important skill for kindergarten students, while number identification skills may also play an important role at the first grade level. Second, these measures are currently commercially available and therefore, educators are likely to use them; and their availability and use, in turn, have implications for many children. Related, AIMSweb®, a website supported by the National Center on Student Progress monitoring, which is sponsored by the U.S. Department of Education Office of Special Education, provides TEN measures. Finally, my pilot study (Petreshock, Codding, Johnson, Russo, & Schaffer, 2006) examined the TEN measures to address a gap in the research at that time. Because that my dissertation was a longitudinal study, using data from my pilot study, I used the same measures across the two years of data collection. Although research provides some important preliminary information on TEN, some crucial research questions remain unanswered. In addition to repeated recommendations to examine the long-term predictive validity of TEN measures over multiple school years, Chard et al. noted that there is a need to examine the reliability of TEN with respect to kindergarten students. To my knowledge, no one has explored the predictive validity of TEN for other teacher-determined academic and behavioral outcomes over multiple school years. Finally, research needs to determine which TEN measures are the best predictors of later mathematics outcomes, such as computation skills.

Pilot Study

In order to begin to address some of the gaps in the TEN literature, I conducted a study to extend the research with a downward extension to kindergarten students

(Petreshock et al., 2006). Specifically, Petreshock et al. investigated whether TEN measures were reliable, valid, and sensitive to growth for kindergarten children. The study tested the hypothesis that TEN measures would demonstrate adequate reliability and validity evidence and I expected that the TEN measures would be sensitive to changes in student performance over time. A total of 82 kindergarten students from 11 classrooms in one suburban public school district located outside the NYC metropolitan area participated in the study. Although research assistants and I assessed 94 students over the course of the study, student absenteeism occurred at both data collection points. Therefore, I analyzed data for only those 82 students present at both data collections points, resulting in the final sample size.

Petreshock et al. recruited participants via a letter sent home to the parents of approximately 200 kindergarten students that explained the purposes of the study. I did not include in the study students whose parents returned forms indicating that they did not wish their children to participate and those students for whom forms parents did not return forms or students who did not provide verbal assent. Of the 82 participants, 35 (43%) were female and 47 (57%) were male. Thirty-three (40%) participants were Caucasian, 26 (32%) were Hispanic, 16 (20%) were African-American/Black and 7 (8%) were Asian. At the final data collection point, participants' ages ranged from approximately 5 years, 6 months to 6 years, 6 months, with an average age of 5 years, 11 months. Five (6%) participants were receiving ESL services and 6 (7%) were receiving other related services such as speech/language therapy or counseling. Seven (8%) of the participants were placed in full-time self-contained special education classrooms. Approximately 44% of the students in the school were eligible for free or reduced lunch.

Petreshock et al. individually administered six measures, four experimental and two criterion, to participants during the Winter (February/March) and Spring (June) of 2006. We created packets containing all measures and counterbalanced the order of experimental measures in attempt to avoid practice effects. The experimental measures were identical to the TEN measures used by Clarke and Shinn (2004) but were modified for kindergarten students to only include numbers to 10, which Clarke and Shinn perceived to be more representative of the kindergarten curriculum requirements. I obtained the measures from <www.aimsweb.com>. The *Woodcock-Johnson-III Applied Problems* (WJ-III-AP) subtest (Woodcock, McGrew, & Mather, 2001) and *M-CBM* (Shinn, 1989) served as criterion measures. I and other school psychology graduate students administered criterion measures in order to determine the concurrent and predictive validity of the experimental measures.

Prior to conducting statistical analyses, I examined both interscorer reliability and procedural integrity by having a second independent observer listen to audio recordings of assessment sessions and record information either onto assessment forms or a procedural integrity form, respectively. With respect to interscorer reliability, a second scorer scored a sample (35%, $N = 57$) of assessment packets from both Winter and Spring 2006, which included all of the measures used in the study and I then examined scores for interscorer agreement. I determined interscorer agreement by dividing the number of agreed upon responses by the number of agreed and disagreed upon responses and multiplying by 100. Interscorer agreement for all of the measures, as a whole, ranged from 97% to 100%, with an average of 99%. An independent observer inspected a sample (32%, $N = 26$) of the Winter 2006 assessment packets using a protocol to examine

procedural integrity. I determined integrity by finding the percentage of correctly implemented assessment procedures out of the total number of assessment procedures. Integrity ranged from 74% to 100% with an average of 93%. It should be noted that the lower bound of integrity (i.e., 74%) was a result of one of the examiners consistently omitting one of the statements indicated in the standardized instructions.

The study used criteria established by Salvia and Ysseldyke (1998) for evaluating reliability in educational contexts in interpreting the results. Accordingly, Salvia and Ysseldyke recommend reliability estimates of .90 or higher for making educational decisions about individual students, reliability estimates of .80 or higher for individual student screening decisions, and reliability estimates of .60 or higher for making decisions about groups of students.

Generally, all four experimental measures exhibited adequate alternate-form reliability, with correlations ranging from .71 to .91. In contrast, an examination of test-retest reliability coefficients indicated that only two of the measures (counting and quantity discrimination measures) yielded acceptable coefficients that for group decision-making. Also, consistent with the findings of Chard et al. (2005), most experimental measures demonstrated reasonable concurrent and predictive validity when used with kindergarten students. Specifically, concurrent validity correlations among the measures and between the measures and a criterion measure ranged from .23 to .61. Predictive validity correlations ranged from .33 to .56. Last, I examined the sensitivity of the TEN measures. It was important to explore the ability of TEN to measure student growth over time, since Clarke and Shinn (2004) designed these measures to monitor progress in a formative evaluation framework. I found sensitivity of the measures to detect change

over time as students improved on all of TEN measures over the 13-week data collection period.

Overall, the pilot study replicated many of Clarke and Shinn's (2004) findings with first grade students. In both studies, a measure of students' ability to discriminate between quantities appeared to have the greatest psychometric support and could be most useful for monitoring progress over time. Chard et al. (2005) also found support for this measure in their study along with a measure of students' ability to identity the missing number in a string of numbers. Inconsistent with my pilot study results, Chard et al. noted that the Number Identification measure should also be considered as a screening tool for first grade students. There were some additional differences among the studies of TEN. For example, contrary to the findings with first grade students (Clarke & Shinn), a measure of students' oral counting skill demonstrated promise for use with kindergarten students. It is not necessarily surprising that the measure of oral counting demonstrated stronger evidence of reliability, validity, and sensitivity to growth for kindergarten students than for first graders, as kindergarten students are at the beginning stages of number sense and some authors (Gersten et al., 2005) have described counting as a prerequisite to other number sense skills. This type of finding has important implications for which TEN measures are most appropriate for students at different grade levels and suggests that the utility of specific indicators of early mathematics abilities varies depending on students' grade level.

Early Numeracy: Extending Technical Adequacy to Contextually Relevant Variables

According to VanDerHeyden et al. (2006), part of the rationale for using CBM is to provide ecologically-valid or contextually relevant assessment in order to avoid

problems associated with measures obtained at one point in time and outside of the classroom. As previously indicated, researchers initially designed CBM to provide teachers with a dynamic assessment tool to guide decisions about student progress relevant to the specific curriculum being used (Deno, 1985). Teacher grades and skill perceptual judgments are an important part of determining the technical adequacy of early measures that are intended for use in the classroom because teacher determined outcomes are often the primary sources of information regarding academic achievement (Cadwell & Jenkins, 1986; Hemingway, Hemingway, Hutchinson, & Kuhn, 1987). These impressions affect daily instructional decision making, student expectations, and teacher-student interactions (Good & Brophy, 1986).

There have been studies of comparisons between teacher ratings of academic competence and achievement measures. For example, Demaray and Elliott (1998) examined the relationship between teacher ratings on the Academic Competence Scale of the Social Skills Rating Scale (SSRS; Gresham & Elliott, 1990) and teachers' item-by-item predictions of performance with students' Kaufman Test of Educational Achievement Brief Form (K-TEA; Kaufman & Kaufman, 1985). They obtained high correlations between teachers' ratings on the Academic Competence Scale of the SSRS and teacher's predictions of performance with students' actual performance on the K-TEA (.70 and .79, respectively). Related to early mathematics CBM in particular, VanDerHeyden and colleagues (2004) found that teacher rankings of preschool student math skills demonstrated moderate to strong correlations with CBMs involving choosing numbers, number naming, counting objects, discriminating among numbers, letters, and shapes and choosing shapes.

To date, research on early numeracy CBM has primarily explored validity by examining relationships among CBM probes and single scores on standardized mathematics measures. Thus, when examining the technical features of early numeracy CBM, it is essential to consider how performance relates to variables important in the context of everyday schooling. Because researchers initially intended the collection of CBM measures to be useful to teachers (Deno, 1985), correspondence with teacher determined measures is critical in order to produce meaningful early mathematics measures that also provide face validity evidence (VanDerHeyden et al., 2004). Contextually-relevant variables might include grade appropriate academic skills, namely computation skills, as well as outcomes that reflect decisions made by classroom teachers such as report card grades as well as teacher ratings of academic skills.

Early Numeracy Indicators of Academic and Behavioral Outcomes

The relationship between academic difficulties and problematic behavior has been well documented (e.g., Hinshaw, 1992). Problem behavior may be conceptualized as externalizing, or under-controlled behaviors such as defiance, aggression, and hyperactivity. Evidence of this relationship is meaningful because problematic behavior in one area may lead to difficulties in the other. Understanding this relationship more clearly may help educators create more effective interventions and safer schools. The notion that TEN performance might predict behavioral difficulties is also of interest in light of recent research (McIntosh, Horner, Chard, Boland, & Good, 2005) indicating that kindergarten reading CBM performance predicts later problematic behaviors in school, as measured by discipline referrals. Specifically, McIntosh et al. found that reading skills measured by CBM at the end of kindergarten predicted the presence of two or more

office disciplinary referrals (ODRs) at the end of fifth grade. What is even more notable is that reading skills in kindergarten were more predictive of later ODRs than were ODRs received in kindergarten. These findings imply that, for some students, academic difficulties precede problem behaviors. McIntosh et al. point out that the importance of this finding lies in the fact that reading difficulties may further place students at risk for behavior problems and hence, risk for non-responsiveness to interventions.

Given the paucity of research on early numeracy measures, it is not surprising that evidence does not exist, at least as far as I am aware, of a relationship between early mathematics CBM performance and behavioral measures. However, some research (Dobbs, Doctoroff, Fisher, & Arnold, 2006) suggests that there is a link between mathematics skills difficulties and problem behaviors in the early years of schooling. Specifically, Dobbs et al. found that initiative, self-control, and attachment were related to better mathematics skills in a sample of preschool children, while overall behavior problems, internalizing symptoms, social problems, and attention problems were related to poorer math skills. Furthermore, the study found that participation in an early math intervention was associated with fewer behavior problems. Examining whether TEN performance in kindergarten predicts problematic school behaviors would add to the research examining the relationship between academics and behavior. In addition, it would extend the existing evidence to support the use of TEN in early identification efforts by possibly further highlighting the importance of remediating academic difficulties early in order to improve behavioral outcomes in addition to academic outcomes.

Visual Quantity Discrimination

With the exception of the oral counting measure, the TEN measures are similar in that they present printed numbers to students. It is possible that performance on measures using this format may be confounded with linguistic skills. Given that investigators (e.g., Gersten et al., 2005) repeatedly cite the ability to use multiple representations of the same number along with making quantity comparisons as important components of number sense; it is somewhat surprising that, as of the inception of this research, previous research lacked measures that required students to distinguish among different types of quantity representations. Therefore, as an addition to the TEN measures, it seems that a measure of visual quantity discrimination, which requires students to make comparison judgments about the magnitude of object sets, is also an important to assess discrimination skills.

Rationale and Hypotheses

Statement of Research Problem

The literature on the psychometric properties of TEN is in its early stages, but provides preliminary support (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Petreshock et al., 2006, etc.) for the use of these measures to assess number sense skills with kindergarten and first grade students. Specifically, research results have demonstrated test-retest, alternate form, and interscorer reliability, and evidence of concurrent and predictive validity of these measures. Although this work represents a good starting point to verify the psychometric properties of early numeracy measures, several important questions remain unanswered.

Most notably, research has not investigated sensitivity and predictive validity of TEN measures across multiple school years. Despite the fact that progress monitoring is a vital aspect of early intervention, Foegen et al. (2007) note that only two studies have examined student growth on early numeracy CBM over time (Clarke & Shinn, 2004; Chard et al., 2005) and that future research needs to continue to address this issue. Several authors have highlighted investigation of TEN over multiple school years as an area of need in the literature to better understand how TEN may be used to screen for early mathematics difficulties and monitor student progress in mathematics in the early years of schooling (e.g., Clarke & Shinn, 2004; Chard et al., 2005; Petreshock et al. 2006). To date, researchers have either considered kindergarten and first grade performance on TEN measures separately or through cross-sectional data. Longitudinal research on TEN performance would help determine the sensitivity of these measures over longer periods of time and could potentially be used to establish guidelines for expected rates of progress across school years. Longitudinal research is also particularly interesting given a recent study demonstrating that brief curriculum-based early literacy skills measures administered in pre-school are significantly predictive of kindergarten and first grade reading outcomes (Missall et al., 2007).

Understanding which kindergarten TEN measures best relate to or predict first grade mathematics outcomes would provide additional types of validity evidence for TEN and help determine which measures serve as the best early indicators of mathematics difficulties. Clarke and Shinn (2004) note that data on the individual contributions of each of the TEN measures has the potential to help educators determine which measures are important to use and whether multiple TEN measures are more

advantageous that a single indicator. In addition, the relationship of TEN to previously described contextually relevant variables, including computation skills, and teacher determined variables such as report card grades, along with teacher evaluations on standardized ratings scales would yield validity evidence that relates to decisions made in the school environment.

Purpose of the Study

I conducted this study in order to address the areas of need indicated in the current research on TEN as stated in the statement of the research problem. Specifically, the purpose of this study was to determine: a) whether TEN measures demonstrate sensitivity to growth from kindergarten to first grade; b) whether TEN performance in kindergarten is correlated with first grade TEN performance; c) whether TEN performance in kindergarten predicts first grade math computation skills (as measured by M-CBM), teacher ratings of mathematics skills, and end of year overall mathematics report card grades; d) whether TEN performance relates to problematic school behaviors, as measured by first grade discipline referrals; and e) to explore the technical features of a VQD measure .

Preliminary research findings indicate that CBM of early mathematics skills demonstrate moderate correlations from preschool to kindergarten (VanDerHeyden et al., 2006). Given these and other findings suggesting that early numeracy measures are sensitive to student growth over one school year and demonstrate adequate validity (Clarke & Shinn, 2004; Chard et al., 2005; Petreshock et al., 2006), I expected that TEN measures would be sensitive to student growth from kindergarten to first grade and would

predict TEN performance and M-CBM performance over multiple school years. Thus, the dissertation examined the following research hypotheses:

H01: Performance on TEN measures will increase significantly from kindergarten to first grade.

H02: Kindergarten TEN performance will be significantly correlated with first grade TEN performance.

H03: Kindergarten TEN performance will be significantly correlated with first grade math computation skills as measured by M-CBM.

H04: Kindergarten TEN performance will significantly predict teacher ratings of mathematics skills.

H05: Kindergarten TEN performance will significantly predict first grade end of year overall mathematics report card grades.

In addition, exploring whether TEN performance in kindergarten predicts later problematic behaviors in school would extend research (e.g., McIntosh et al., 2005) on the relationship between academic and behavioral difficulties. Specifically, it was of interest to explore whether a relationship between kindergarten mathematics skills and behavior problems similar to the relationship between kindergarten reading skills and behavior problems found by McIntosh et al. 2005 exists. Given findings suggesting a link between early math skills and problematic behaviors (Dobbs et al., 2006), I expected that kindergarten TEN performance would predict behavior problems in the subsequent school year, as measured by first grade disciplinary referrals. Therefore, the dissertation tested the following hypothesis:

H06: Kindergarten TEN performance will significantly predict disciplinary referrals in the first grade.

Given that the ability to use multiple representations of the same number and compare quantities are accepted components of number sense (Gersten et al., 2005); the ability to make judgments about pictorially, rather than symbolically, represented numbers may be an additional important number sense skill to assess. Since visual quantity discrimination skills are part of the total construct of number sense, it would be expected that they are related to other number sense skills. Therefore, the dissertation tested the following hypothesis:

H07: First grade performance on a Visual Quantity Discrimination (VQD) measure will be correlated with TEN performance and first grade outcomes.

CHAPTER III

Methodology

This chapter provides an explanation of the participant selection and sample characteristics, as well as descriptions of the setting, measures, research design, and data collection procedures. In addition, the chapter explains interscorer agreement and treatment integrity methods. Finally, the chapter concludes with a presentation of the data analysis procedures.

Participant Selection

A total of 61 first grade students from 13 classrooms in three schools of a suburban public school district located outside the New York City metropolitan area participated in the study. I recruited the dissertation sample from the participants in my pilot study because the dissertation covers student performance over two academic years (kindergarten and first grade). Specifically, the pilot study provided the kindergarten data for the dissertation. Although I obtained complete kindergarten data from just 82 students, a larger pool of students ($N = 94$, out of approximately 200 solicited, 47% acceptance rate) participated (i.e., provided some data) in some phase of the pilot study. (Appendix A provides the parental consent form for the pilot study.)

To complete the dissertation, therefore, I sought to recruit as many students from the initial pilot pool of 94 as possible to obtain first grade data to compare with the kindergarten data collected during the pilot. Of the 94 potential participants, 14 students had moved out of the district from kindergarten to first grade, leaving just 80 possible participants for first grade assessment.

I gave consent forms explaining the purpose of the study, including benefits and risks, and inviting continued participation in the study (see Appendix B) to teachers to

give to the parents of the 80 students. Teachers initially distributed forms to parents during parent-teacher conferences. If parents were not in attendance at conferences, teachers sent consent forms home in students' folders. This initial distribution resulted in parents returning 43 forms. In order to improve the response rate, teachers conducted a second distribution to parents. This time, teachers sent forms to all parents who had not returned forms, not just parents who did not attend parent teacher conferences, in students' folders. This second mailing yielded 15 additional participants, which brought the number of participants to 58. In a final attempt to recruit as many participants as possible, teachers directly contacted parents who still had not returned forms to request that they return the forms. This procedure resulted in 10 additional forms returned for a total of 68 forms returned after the third attempt, constituting a final return rate of 85% (68 of 80) of forms from parents solicited. I did not include in the study students whose parents returned forms indicating that they did not wish their children to participate and students for whom parents did not return forms. Out of the 68 forms returned, 61 granted permission for students to participate (90% of forms returned). All 61 students with parental consent to participate provided verbal assent. The sample was large enough to meet requirements for statistical power as provided by Cohen (1992). Specifically, given a large effect size at the .05 level, there should be a sample of $N = 28$ for correlational analyses and a sample of $N = 38$ for multiple regression analyses with four predictor variables (Cohen, 1992).

To summarize, only 80 of the 94 students who participated in the pilot study (and thus provided data for the kindergarten year of this dissertation) still lived in the district when I collected first grade data to complete the dissertation. Parents of 61 of these 80

students provided consent for participation in the current study. Thus, 65% (61 of 94) of students who participated in at least some part of the pilot (kindergarten) also participated in the dissertation follow-up (first grade). Also, of the 200 students that I solicited when they were in kindergarten, 30.5% participated in all four assessments.

Comparison of demographic characteristics of study completers and dropouts. I used chi-square analyses to determine if there were any significant differences in the demographic characteristics of students who moved or did not receive consent to continue participation in the second (first grade) year (attrition group, $n = 31$) and those who participated over the two year period (final sample, $n = 61$) of the study (see Table 1 for participant characteristics for final sample). Appendix C provides characteristics of students in the attrition group. The percentage of students in the attrition group and the final sample did not differ by gender, $\chi^2 (1, N = 92) = 0.26$, $p > .05$, receipt of ESL services, $\chi^2 (1, N = 92) = 1.64$, $p > .05$, or receipt of other related services, $\chi^2 (1, N = 92) = 1.28$, $p > .05$. However, the ethnic groups of the attrition group and final sample differed significantly, $\chi^2 (1, N = 92) = 8.11$, $p < .05$. The attrition group consisted of a greater proportion of Hispanic students ($n = 16$, 48.5%) than the final sample (see Table 1), while the proportion of African-American students ($n = 2$, 6.1%) was smaller in the attrition group than in the final sample (see Table 1). The percentages of Asians and Caucasians were relatively similar in both groups. This finding may reflect greater transience among the Hispanic population in the district along with a more lasting African-American community, at least during the two years of the study. In addition, the attrition group and final sample differed significantly in the proportion of students receiving special education services, $\chi^2 (1, N = 92) = 4.45$, $p < .05$. There were no

students receiving special education services in the attrition group, indicating that students with special education designations all remained in the district and received consent for participation over both years.

Participants

Table 1 presents the characteristics of the final sample of 61 students who participated in the dissertation by providing both kindergarten and first grade data.

Table 1

Participant Demographics

Variable	*N*	Percent of Sample
Gender		
Male	34	55.7%
Female	27	44.3%
Ethnicity		
African-American	15	24.6%
Asian	5	8.2%
Caucasian	27	44.3%
Hispanic	14	23.0%
ESL services		
Yes	2	3.3%
No	59	96.7%
Special education services		
Yes	8	13.1%
No	53	86.9%
Related services		
Yes	6	9.8%
No	55	90.2%

I compared the demographic characteristics of the final sample to the overall demographic characteristics of the district according to the most recently available New York State district report card data. Specifically, according to the New York State Education Department (NYSED, 2006), 19% of the district population of students are African-American, 29% are Hispanic, 6% are Asian, and 45% are Caucasian. Additionally, 9% of the district population is classified as Limited English Proficient (LEP). Thus, while the study participants approximate the district population in terms of ethnicity, LEP students appear to be underrepresented in the sample based on the percentage of participants receiving ESL services.

Setting

I conducted the study in three elementary schools in a public school district located in a suburb of New York City. There were a total of 13 first grade classrooms in the district with 5 first grade classrooms in one school, and 4 first grade classrooms in each of the other two schools. Students from all 13 first grade classrooms were involved in the study. The average elementary class size for the district was 21 students (NYSED, 2006). The number of students from each class who participated ranged from 1 to 10, with an average of 5 participating students per classroom. Appendix D provides complete data on the number of students from each classroom.

Data collection took place in hallways outside student classrooms. The hallways were spacious, clean, bright, and often adorned with student work. I used this location for several reasons: it allowed for ease of transitioning students to the research situation, minimized time spent in transition, and allowed students to remain in view of school staff. In order to minimize disruptions, I attempted to conduct the study when classes

were not changing and paused data collection during times of busy hallway traffic (e.g., between class periods).

Experimental Measures

The experimental measures consisted of four Tests of Early Numeracy (TEN) measures: (a) Oral Counting (OC), (b) Number Identification (NI), (c) Quantity Discrimination (QD), (d) Missing Number (MN), (Clarke & Shinn, 2002; Edformation (TEN), 2002). I created the fifth experimental measure, the Visual Quantity Discrimination (VQD) measure, for this study in order to explore other possible skills related to number sense. With the exception of the VQD measure, I obtained all experimental measures from <ww.aimsweb.com>. (Readers should note the extensive review of TEN research in Chapter II beginning on page 41.) I created the VQD measure and Appendix E presents this measure. The TEN measures assess skills addressed in the New York State Learning Standards for Mathematics (e.g., understanding numbers and the relationship between numbers), which are the basis for the local curriculum (Math Core, 2006).

I initially emailed Mr. Gary Germann, former owner of the Edformation, which publishes the TEN measures to obtain permission to use them. Mr. Germann indicated that he sold Edformation to Harcourt Assessment and provided the appropriate new contact information. Subsequently, Mr. Jay Anderson of Harcourt Assessment and AIMSweb® stated that I could obtain the measures through a subscription to AIMSweb®. Appendix F contains the relevant correspondence. A more detailed description of each measure follows.

Oral Counting (OC). The OC measure requires students to count orally from 1 as

the experimenter records student responses on a scoring sheet. If a student struggles or hesitates for more than 3 seconds, the examiner instructs him or her to say the next number. Examiners report performance as the amount of numbers correctly counted in 1 minute. Possible scores range from 0 to over 100, depending on how many numbers the student correctly counts in 1 minute. Test-retest reliability correlations for the OC measure range from .68 to .80, while the alternate-form reliability is reported as .93 and inter-scorer reliability is .99 (Clarke & Shinn, 2004; Petreshock et al., 2006). Concurrent validity correlations of the OC measure with The Woodcock Johnson-III Applied Problems subtest (Woodcock, McGrew, & Mather, 2001) range from .37 to .64. Concurrent validity correlations with math computation, as measured by M-CBM, range from .49 to .50 (Clarke & Shinn, 2004). Studies also found this measure to be sensitive to growth for kindergarten and first grade students (Clarke & Shinn, 2004; Petreshock et al., 2006).

Number Identification (NI). The NI measure requires students to orally identify numbers when presented with printed number symbols. At the kindergarten level, students identify numbers from 0 to 10. At the first grade level, students identify numbers from 0 to 20. Examiners give students a sheet of random numbers in an 8 x 7 grid and require them to name the numbers presented on the sheet. There are a total of 56 items on the sheet. If a student struggles or hesitates for more than 3 seconds, the examiner instructs him or her to identify the next number. Examiners report performance as the number of numbers correctly identified in 1 minute. If a student completes the task in less than 1 minute, the examiner prorates his or her score. Thus, the possible range of scores is from 0 to 56 or more. Test-retest reliability correlations of the NI measure range from

.46 to .85, while alternate-form reliability correlations are reported as ranging from .71 to .93 and inter-scorer reliability is .99 (Clarke & Shinn, 2004; Petreshock et al., 2006). Concurrent validity correlations of the NI measure with The Woodcock Johnson-III Applied Problems subtest (Woodcock, McGrew, & Mather, 2001) range from .23 to .66 (Clarke & Shinn, 2004; Petreshock et al.). Concurrent validity correlations with math computation, as measured by M-CBM, range from .60 to .66 (Clarke & Shinn, 2004). Studies also found NI to be sensitive to growth for kindergarten and first grade students (Clarke & Shinn, 2004; Petreshock et al.).

Quantity Discrimination (QD). The QD measure requires students to name which of two visually presented numbers is larger. Experimenters give participants a grid with 28 boxes containing random numbers, constituting a total of 28 items. At the kindergarten level, measures contain numbers from 0 to 10. At the first grade level, measures contain numbers up to 20. One number is always larger than the other. If the participants stop, struggle, or hesitate for more than 3 seconds, examiners encourage them to try the next one. Examiners report performance as the number of correctly identified larger numbers in 1 minute. If a student completes the task in less than 1 minute, the examiner prorates his or her score. Thus, the possible range of scores is from 0 to 28 or more. Test-retest reliability correlations of the QD measure range from .65 to .85, while reports of alternate-form reliability correlations range from .89 to .92 and inter-scorer reliability is .99 (Clarke & Shinn, 2004; Petreshock et al., 2006). Concurrent validity correlations of the QD measure with a standardized measure of mathematics problem solving (The Woodcock Johnson-III Applied Problems subtest; Woodcock, McGrew, & Mather, 2001) range from .57 to .79 (Clarke & Shinn, 2004; Petreshock et

al.). Concurrent validity correlations with math computation, as measured by M-CBM, range from .71 to .79 (Clarke & Shinn, 2004). QD is also sensitive to growth for kindergarten and first grade students (Clarke & Shinn, 2004; Petreshock et al.).

Missing Number (MN). The MN measure requires students to name the missing number from a string of numbers. Kindergarten measures contain numbers from 0 to 10 while, first grade measures contain numbers from 0 to 20. Examiners give students a sheet with 21 boxes on it, each counting as one item and each containing a string of 3 numbers. One number is missing from the string. If a student struggles or hesitates for 3 seconds or more, the examiner encourages him or her to try the next one. Examiners report performance as the number of correctly identified missing numbers in 1 minute. If a student completes the task in less than 1 minute, the examiner prorates his or her score. Thus, the possible range of scores is from 0 to 21 or more. Test-retest reliability correlations of the MN measure range from .53 to .79 while studies report alternate-form reliability correlations ranging from .78 to .86 and inter-scorer reliability is .98 (Clarke & Shinn, 2004; Petreshock et al., 2006). Concurrent validity correlations of the NI measure with the Woodcock Johnson-III Applied Problems subtest (Woodcock, McGrew, & Mather, 2001) range from .37 to .69 (Clarke & Shinn, 2004; Petreshock et al.). Concurrent validity correlations with math computation, as measured by M-CBM, range from .71 to .75 (Clarke & Shinn, 2004). MN was also found to be sensitive to growth for kindergarten and first grade students (Clarke & Shinn, 2004; Petreshock et al.).

Visual Quantity Discrimination (VQD). The VQD measure requires students to point to which of two visually presented groups of objects is larger. I created the VQD measure due to concerns that all of the other measures used printed number symbols and

thus, may be confounded with linguistic skills. Although Gersten et al. (2005) highlighted the ability to discriminate between quantities as one of the key components of number sense, at the inception of this study, previous research had not examined measures that required students to distinguish among pictorially (as opposed to symbolically) represented quantities. Floyd et al. (2006) have since examined a somewhat similar measure in their study with preschool students. However, there is no measure of this type for kindergarten students. Examiners give participants a grid with 40 boxes, each constituting one item, containing groups of circles of differing quantities up to 20. One set of circles is always larger than the other. If the participant stops, struggles, or hesitates for more than 3 seconds, the examiners encourages him or her to try the next group of circles. Examiners report performance as the number of correctly identified larger sets of circles in 1 minute. If a student completes the task in less than 1 minute, the examiner prorates his or her score. Thus, the possible range of scores is from 0 to 40 or more. Given that VQD is a newly developed measure created to explore other skills hypothesized to be related to number sense, psychometric data were not available prior to its use. I conducted reliability and validity analyses using data from the current study and present them fully in the results section.

Criterion Measures

Mathematics-Curriculum-Based Measurement (M-CBM) (Shinn, 1989), office discipline referrals (ODRs), final overall math report card grades, and the Total Mathematics Score on the Academic Competency Evaluation Scales (ACES; DiPerna & Elliot, 2000) served as criterion measures. I used criterion measures to determine the longer-term predictive validity of the experimental measures and the relationship between

the experimental measures and criterion measures.

M-CBM. I obtained grade 1 CBM computational probes from <www.aimsweb.com> (Edformation, 2002). Examiners administered three, two-minute M-CBM probes created for the middle and end, or for the winter and spring, administration times respectively, of grade 1. The probes require students to solve addition and subtraction problems with one or two-digit numbers. Examiners report performance as the number of digits correct per minute. Data analyses used students' median scores among the three probes. Foegen, Jiban and Deno (2007) reviewed 17 research studies on progress monitoring measures for elementary mathematics (i.e., M-CBM). Reliability coefficients generally exceeded .80 with test-retest reliability estimates greater than alternate-form reliability estimates. According to Thurber, Shinn, and Smolkowski (2002), interscorer reliability is .83. These authors also examined the relationship between M-CBM (i.e., computation) probes with math applications measures that assessed students' ability to apply basic math facts and principles. They found that computation and applications were separate, but highly related, skills with M-CBM being a measure of computation. Criterion related validity correlations among basic facts CBM probes and other measures of computation range from .30 to .83 (Foegen, Jiban, & Deno, 2007; Thurber, Shinn, & Smolkowski, 2002).

Office disciplinary referrals (*ODRs).* I reviewed participating students' educational records with participating school staff at the end of first grade in order to determine the number of ODRs each student received over the school year. I recorded the number of ODRs for each participant. Irvin, Tobin, Sprague, Sugai, and Vincent (2004) provide a review of the literature on ODRs to establish their validity as indicators of

school-wide behavioral climate. These authors cite moderate negative relationships with academic achievement and commitment to schooling, and moderate to strong positive relationships with rebellious, antisocial, and other problematic behaviors. In another study, Walker, Cheney, Stage, and Blum (2005) found that students with multiple ODRs scored higher than one standard deviation above the mean on a problem behavior subscale of a standardized behavior rating scale.

Final overall math report card grades. At the end of the first grade school year, I obtained all students' overall mathematics report grade grades. In the participating district, students receive grades ranging from 1 to 4. A grade of 1 indicates that the student's performance is below grade level expectations, a 2 indicates that the student meets grade level expectations with support, a 3 indicates that the student performs on grade level, and a 4 indicates that the student performs above grade level expectations. I obtained report card grades directly from teachers.

ACES Mathematics total score. At the end of the first grade school year, teachers rated participants on the ACES mathematics scale (ACES-M) using a 5 point likert scale where ratings ranged from 1 *(far below)* to 5 (*far above*). The ACES is a kindergarten through twelfth grade rating scale with 73 items that evaluates academic problems and targets areas for intervention, such as specific academic skill areas and behaviors that contribute to academic success. More specifically, the ACES yields Academic Skills and Academic Enablers composite scores. The Academic Skills composite consists of Reading/Language Arts, Mathematics, and Critical Thinking scales. The Academic Enablers composite is comprised of the Interpersonal Skills, Engagement, Motivation, and Study Skills scales. The ACES has a reported internal consistency reliability

coefficient of .94 to .99 and test-retest reliability ranges from .88 to .97 (DiPerna & Elliot, 1999). Evidence for the validity of the ACES was based on comparisons to other measures of academics including the IOWA Basic Skills Composite and the Academic Competence scales of the Social Skills Rating System. Validity correlations range from .66 to .80 (DiPerna & Elliot, 1999). The Mathematics composite of the ACES contains 8 items and takes approximately 5 minutes or less to complete for each student. Items in the mathematics composite ask teachers to rate students' skill level in computation, pattern analysis, measurement, understanding of spatial relationships, mental math, using numbers to solve daily problems, breaking down complex problems, and problem-solving. Individual item scores combine to form the Mathematics Total Score. Scores range from 8 – 40, with higher scores indicating greater student math proficiency. I used this score in statistical analyses examining the relationship between the TEN measures and ACES-M ratings.

Research Design

This study employed a longitudinal design and measured student performance over two years. Specifically, the study measured student performance from kindergarten (year 1) through first grade (year 2). During year 1 (kindergarten), examiners individually administered participants each of the TEN measures twice, once during the winter and once during the spring of the year, approximately 13 weeks apart. Examiners also administered the M-CBM in the spring of the kindergarten school year. During year 2 (first grade), examiners individually administered participants each of the TEN measures, the VQD measure, and M-CBM during the winter and spring of the school year, approximately 13 weeks apart. I also collected grades, ODRs, and ACES scores at the

end of the first grade year. Kindergarten TEN performance served as the predictor variables while first grade TEN and VQD performances, first grade winter and spring M-CBM performance, end of year overall mathematics report card grades, teacher ratings of mathematics skills, and first grade disciplinary referrals were the outcome variables.

Order of Administration of Measures

I created packets containing the five experimental measures: (a) OC, (b) NI, (c) QD, (d), MN, and (e) VQD. Each packet also contained the M-CBM criterion measure for examiners to give as the final measure. I counterbalanced the order of the experimental measures in an attempt to avoid practice or fatigue effects. To achieve counterbalancing, I staggered the order of presentation of the experimental measures in the assessment packets so that, while each packet contained identical measures, the order in which examiners would administer them to students varied. Specifically, counterbalancing produced five combinations: abcde, bcdea, cdeab, deabe, eabcd. Readers will note that packets contained only six of the nine measures used in this dissertation. This is because I obtained the remaining three criterion measures (grades, ODRs, and ACES-M scores) either from students' first grade records or from their first grade teachers. I randomly distributed the packets to the examiners.

Procedures

Administration of mathematics measures. Prior to beginning all research procedures, I obtained district permission (Appendix G) and Institutional Review Board (Appendix H) approval to conduct research. During the kindergarten year after they obtained consent and assent, examiners administered the four experimental measures to students during Winter (February/March) 2006 and Spring (June) 2006. In the spring,

examiners also administered M-CBM in addition to the four measures. Then, during the first grade year, after parents gave consent and students provided assent, examiners individually administered experimental and M-CBM measures to students at two points in the school year, Winter (February/March) and Spring (June) 2007. In the first grade year, data collection sessions occurred approximately 13 weeks apart. There were approximately 39 weeks between the spring kindergarten testing and the winter first grade testing. I selected this time frame for data collection to correspond to the data collection schedule in my pilot study on the basis of collection procedures used in previous studies of TEN (Clarke & Shinn, 2004; Chard et al., 2005). Examiners administered all experimental measures according to standardized procedures (Clarke & Shinn, 2002). Individual testing sessions lasted approximately 20 minutes per participant at each data collection point. Upon completion of the measures at each data collection point, examiners verbally thanked participants for their participation and provided each of them with a sticker.

Examiner training. Examiners included the author and seven school psychology graduate students. Four of the examiners, including the author, were doctoral level students and the remaining four were master's level students. I trained examiners in test administration procedures via participation in a one-hour training session. During this session, I provided examiners with test materials, reviewed test protocols and standardized instructions, and gave them the opportunity to practice the procedures and ask questions. In addition to this training, CBM training was part of the masters or doctoral coursework of all of the examiners. Although one examiner did not have formal classes that addressed CBM, her dissertation advisor supervised this examiner in self-

directed study on CBM. Two of the examiners reported having taken one class that addressed CBM and the remaining examiners, including the author, reported that they had taken two more classes that addressed CBM, at least one of which required that they conduct curriculum-based assessment/measurement.

Collection of additional criterion measures. I collected the remaining first grade data - report card grades, grade retention and discipline referral information, and teacher responses on the mathematics portion of the ACES. I obtained report card grades directly from teachers at the end of the school year. I obtained grade retention and discipline referral data from various sources at each school including an administrator, a social worker, and a school psychologist though direct, informal interviews. Finally, I obtained teacher ratings on the mathematics portion of the ACES from teachers. I distributed a letter and sufficient blank copies of the ACES to the 13 teachers of the participating students' classrooms. The letter explained the purpose of collecting such ratings, assured teachers that ratings were confidential, and provided written directions. I also provided verbal directions at the time that I distributed the forms. I gave teachers the option to return the forms directly to me during data collection days or to return them by mail with an addressed and stamped envelope that I provided. Out of the 13 teachers, 10 returned forms that resulted in ACES ratings for 38 students. The remaining 23 students in the sample were in the three classrooms whose teachers did not return forms. I made subsequent requests for the teachers to return forms, but this did not produce additional forms. Thus, teachers provided ACES for just 62% of the sample. Although this return number is notably smaller than the total sample, it does meet requirements for statistical power, considering a large effect, according to Cohen (1992).

Analysis Procedures

I used several data analysis procedures in this study: (a) interscorer agreement and procedural integrity, (b) descriptive statistics, (c) repeated measures ANOVAs, (d) Pearson product moment correlations, and (e) multiple and logistic regression.

Interscorer agreement and procedural integrity. First, it was critical to ensure that results were not due to incorrect or inconsistent scoring or administration of the experimental measures. Therefore, in order to determine whether examiners rated all participants consistently on the experimental measures and exposed participants to the same standardized procedures, I examined a sample (30%) of assessment protocols for interscorer agreement and assessment protocol integrity. To determine scoring and procedural integrity, it is customary for researchers to examine a sample of 30% of protocols (e.g., Clarke & Shinn, 2004). I determined interscorer agreement by having an outside scorer re-score 30% of the assessment protocols and then compared the outside scorer's scores with those of the examiner to obtain agreement and disagreement between these two people.

Three of the examiners were trained in CBM scoring procedures and therefore scored the packets they administered. I served as the outside scorer for these packets. I served as the initial scorer on the remaining packets, some of which I administered. I trained another graduate student, who is also a teacher, in CBM scoring procedures and she served as the outside scorer for packets where I was the initial scorer. Thus, for all of the packets used to determine interscorer agreement, I either served as the initial or second (outside) scorer. After obtaining the independent scores from both scorers, I divided the number of agreed upon responses by the number of agreed and disagreed

upon responses and multiplied the dividend by 100 [Agreements/(Agreements + Disagreements) x 100].

In order to determine assessment protocol integrity, an outside observer reviewed audio recordings of 30% of the testing sessions with the use of a protocol detailing assessment procedures (see Appendix I), and noted the assessment procedures that examiners correctly implemented. I served as the outside observer for testing sessions where I was not also the examiner. The aforementioned outsider scorer for interscorer reliability, who was not involved with data collection, served as the outside observer for testing sessions where I was the examiner. I reported treatment integrity as the percentage of correctly implemented assessment procedures out of the total number of assessment procedures. Although it may have been preferable to have an observer directly observe testing sessions rather than to use audio recordings; since examiners gave all directions orally, and students responded orally for four out of the five experimental measures, I believed that having one observer rate audio recordings would yield valid treatment integrity percentages.

I or the outside observer checked at least one testing session for each examiner. However, the number of sessions recorded, and therefore checked, varied for each examiner for several reasons. For one, many participants did not provide consent for audio-recording. Since students were paired with examiners randomly (on the basis of timing and availability on testing days), some examiners happened to be paired with more students whom parents allowed to be recorded than were other examiners. In addition, due to personal availability, some examiners conducted more assessment session than others, and therefore, provided more recorded sessions. Finally, some of the examiners

experienced technical difficulties with recording devices and therefore, I could not conduct accurate checks of such testing sessions.

Descriptive Statistics. Means and standard deviations described participant characteristics and overall performance on each of the measures.

Repeated measures ANOVAs. I measured the sensitivity of TEN measures from kindergarten to first grade, or whether students made significant progress from kindergarten to first grade, with repeated measures ANOVAs in a series of analyses, one for each measure. Specific measure performance at each data point (i.e., Winter 2007, Spring 2007, etc.) was the within-subjects factor, and the ANOVA tested for statistically significant change in performance over time. In order to describe student progress on each of the measures, I determined growth in digits per week by computing the difference between scores and dividing by the number of weeks between assessments (Clark & Shinn, 2004). For example, in order to determine growth from the winter of first grade to the spring of first grade, for each measure, I found the difference between the average winter 2007 score and the average spring 2007 score and divided by 13, the number of weeks between these testing times.

Correlations and regression equations. In order to answer the research questions regarding the predictive validity of kindergarten TEN performance for first grade TEN and M-CBM performance, I calculated Pearson product moment correlations for all the data as an initial step. In addition, I conducted 12 regression analyses in order to explore the relationship between kindergarten TEN performance, first grade M-CBM performance, ACES scores, and report card grades in further depth. First, four regression analyses examined kindergarten performance on each TEN measure (predictor variables)

where M-CBM performance was the outcome variable as follows: (a) kindergarten winter TEN as predictors of first grade winter M-CBM, (b) kindergarten winter TEN as predictors of first grade spring M-CBM, (c) kindergarten spring TEN as a predictors of first grade winter M-CBM, and (d) kindergarten spring TEN as predictors of first grade spring M-CBM. Second, I used multiple regression analysis to examine performance on each of the kindergarten TEN measures as the predictor variable with the ACES-M score as the dependent variable. For multiple regression analyses, I initially entered all variables into the equation in order to evaluate the overall model. Then, I used backwards deletion procedures to evaluate the contributions of individual variables, where I successively removed predictors on the basis of *B* significance levels. I used the standard criterion of $p = .10$ for inclusion in the model (McIntosh et al., 2006). I selected this approach because of the exploratory nature of the study and because I did not know which variables would be the best predictors (McIntosh et al; Cohen & Cohen, 1975).

Third, I used ordinal regression analysis to examine the predictive validity of the kindergarten TEN measures for students' first grade end of year overall mathematics report card grade. I used this approach since the dependent variable, report card grades, was an ordinal variable (Norusis, 2008). In addition, ordinal regression can examine the probability of obtaining a particular score or less. Since descriptive statistics demonstrated that students who had grades of 3 and 4 performed very similarly on the experimental measures, it was of interest to compare collectively these students to those who were given a grade of 2. Thus, in these equations the outcome was the probability of obtaining a grade of 2 or a grade of 3 or higher. One of these equations used the kindergarten winter TEN measures as the predictors and the other used the kindergarten

spring TEN measures as the predictor variables.

CHAPTER IV

Results

This chapter presents the results of statistical analyses. First, it presents results regarding interscorer agreement and procedural integrity. The chapter then provides descriptive statistics for experimental measures as well as M-CBM and ACES-M. Third, the chapter reports results of the ANOVAs performed to examine the sensitivity of the measures. Fourth, the chapter presents results regarding the validity of the experimental measures based on both correlation and regression analyses. Last, the chapter gives findings for the Visual Quantity Discrimination measure.

Interscorer Agreement and Procedural Integrity

According to procedures detailed in Chapter III, I assessed both interscorer agreement and procedural integrity by having an independent observer listen to audio recordings of assessment sessions and record information either onto assessment forms or onto procedural integrity forms, respectively. With respect to interscorer agreement, a second scorer scored a sample (30%, $n = 37$) of assessment packets from both winter and spring 2007 first grade assessments, which included the five experimental measures used in the study and M-CBM, and I then examined the two scores for interscorer agreement, by dividing the number of agreed upon responses by the number of agreed and disagreed upon responses and multiplying by 100. Interscorer agreement for all of the measures, as a whole, ranged from 93.4% to 100%, with an average of 99.4%. Appendix J provides interscorer agreement data for each individual test packet. The high interscorer agreement ratings are not surprising given that assessment procedures for all measures were standardized.

An independent observer examined a sample (30%, $n = 37$) of the winter and spring 2007 assessment packets using a protocol to examine procedural integrity (see Appendix I). I determined treatment integrity by finding the percentage of correctly implemented assessment procedures out of the total number of assessment procedures. The observer rated assessment procedures that included correct reading of standardized directions and correct timing of test administration. Integrity ranged from 66.7% to 100% with an average of 97.4%. It should be noted that the lower bound of integrity (i.e., 66.7%) was due to an examiner's failure to read standardized directions accurately and occurred for only one participant. Despite this break in standardization, this particular participant's data do not appear to be invalid as the examiner explained the assessment task to the participant in general terms (instead of verbatim), offered the participant the opportunity to ask questions if he did not understand the task, and the participant was able complete the task. It should be noted that the procedure for collecting interobserver agreement and procedural integrity data was identical for the kindergarten data from winter and spring 2006. I reported these results in my pilot study and described them in Chapter II (p. 52).

Descriptive Statistics

Table 2 displays the means and standard deviations for the experimental measures, M-CBM, and ACES-M at kindergarten and first grade data collection points. From visual inspection with respect to the TEN measures, these data illustrate that in both Winter and Spring 2006, kindergarten students scored highest on OC followed by NI, QD, and MN measures, respectively. The same pattern emerged in the first grade. Students' performance on the additional VQD measure fell between their performance on

the QD and MN measures in Winter 2007 and between the NI and QD measures in Spring 2007. In addition, students' performance was variable on each of the measures over both school years, indicating a wide range in students' skills. Lastly, it is clear that students consistently improved on all four TEN experimental measures from the first data collection point in kindergarten (i.e., Winter 2006) to final data collection point in first grade (i.e., Spring 2007). In the next section, I tested the statistical significance of many of these differences.

Table 2

Average Kindergarten and First Grade Performance on Experimental and Criterion Measures

	Kindergarten		First Grade	
	Winter 2006	Spring 2006	Winter 2007	Spring 2007
	(N = 57)	*(N = 55)*	*(N = 61)*	*(N = 61)*
Measures	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Experimental				
OC	50.30 (17.09)	57.95 (16.23)	74.38 (14.84)	78.73 (14.84)
NI	48.13 (14.41)	50.65 (16.65)	55.00 (13.68)	56.93 (12.47)
QD	17.61 (12.77)	20.73 (12.84)	31.41 (9.53)	34.18 (8.67)
MN	8.79 (6.70)	12.97 (5.71)	18.69 (7.67)	18.98 (6.03)
VQD	- -	- -	29.05 (7.24)	34.34 (10.02)
Criterion				
M-CBM	- -	1.76 (2.24)	13.72 (5.65)	11.82 (6.31)
ACES	- -	- -	- -	22.82 (5.69)

Note. OC = oral counting; NI = number identification; QD = quantity discrimination; MN = missing number; VQD =visual quantity discrimination.

Sensitivity

Since Clarke and Shinn (2002) designed the TEN measures to monitor student progress, I examined the sensitivity of each measure, or the ability of the measure to detect significant changes in students' scores over time. Students' scores increased on all of the experimental measures at each data collection point from kindergarten to first grade. I examined growth in digits per week by determining the difference between scores and dividing by the number of weeks between assessments (Clarke & Shinn, 2004). Within each school year, there were 13 weeks between the two assessment points. There were total of 68 weeks between the first assessment in kindergarten (i.e., Winter 2006) and the last assessment in first grade (i.e., Spring, 2007). From the first data collection point in kindergarten to the final data collection point in first grade, on average, students' scores on the OC measure increased by 28.42 digits (0.43 digits per week). Scores on the QD, MN, and NI measures increased by an average of, 16.3 digits (0.24 digits per week), 10.4 digits (0.15 digits per week), and 8.5 digits (0.12 digits per week), respectively.

When I considered kindergarten and first grade performance separately, the pattern was somewhat similar. For the kindergarten measures, students showed the most growth on the OC (0.65 digits per week) measure followed by the MN (0.33 digits per week), QD (0.21 digits per week), and NI (0. 11 digits per week) measures, respectively. For first grade measures, students showed the most growth on the VQD measure followed by the OC, QD, NI, and MN measures, respectively. Table 3 reports data on growth for each measure.

Using the scores from kindergarten winter, kindergarten spring, first grade winter,

and first grade spring data points, I conducted five repeated measures ANOVAs to examine whether growth from kindergarten through first grade for each measure was statistically significant. For TEN measures, I entered scores from all four testings (two in kindergarten and two in first grade) into the analysis. For the OC (F [3, 48] = 54.97, $p <$ .001, $\eta^2 = .78$, a large effect size), NI (F [3, 48] = 5.471, $p = .003$, $\eta^2 = .26$, a large effect size), QD (F [3, 48] = 31.01, $p <.001$, $\eta^2 = .66$, a large effect size), and MN (F [3, 48] = 60.37, $p < .001$, $\eta^2 = .79$, a large effect size) measures, growth was statistically significant and not due to chance. I only administered the VQD measure during first grade, so I entered the scores from the winter and spring testings into the analysis. For the VQD measure, growth was also not due to chance, F (1, 60) = 19.21, p < .001 ($\eta^2 = .24$, a large effect size).

I conducted paired *t*-tests as post hoc analyses in order to examine the difference between testing pairs (i.e., kindergarten winter and kindergarten spring; kindergarten spring and first grade winter; first grade winter and first grade spring). In order to control for the possibility of making a Type I error, for each measure, I divided .05 by the number of comparisons (.05/3), resulting in a significance level of $p < .02$. Six of the 12 pairs were statistically significant at the $p < .02$ level. First, Table 4 illustrates that students' growth on OC was significant within kindergarten and between kindergarten and first grade, but not within first grade. Second, growth on the NI measure between testing time pairs was not significant. Third, growth on QD was not significant within kindergarten but was significant from kindergarten to first grade and within first grade. Fourth, growth on MN was significant within kindergarten and from kindergarten to first grade, but not within first grade. Still, the results of the ANOVAs measuring overall

growth supported the hypothesis (H01) that students would make significant progress on TEN measures from kindergarten to first grade.

Table 3

Student Growth on TEN and VQD Measures During Kindergarten, First Grade, and Across Both Grades

Measure	Growth in Units Per Week		
	Kindergarten (13 weeks)	First Grade (13 weeks)	Total (68 weeks)
	M (SD)	*M (SD)*	*M(SD)*
Oral Counting	.65 (1.09)	. 34 (1.24)	.43 (.26)
Number Identification	.11 (1.30)	.15 (1.02)	.12 (.22)
Quantity Discrimination	.21 (.71)	.21 (.64)	.24 (.18)
Missing Number	.33 (51)	.02 (.50)	.15 (.10)
Visual Quantity Discrimination	-	.41 (.73)	-

Table 4

Post Hoc Analyses of Differences between Testing Pairs

Pair	*t*	*df*	Sig.
OC K winter - OC K spring	-4.27	50	< .001
OC K spring -OC 1 winter	-7.48	54	< .001
OC 1 winter - OC 1 spring	-2.12	60	.038
NI K winter - NI K spring	-.61	50	.546
NI K spring - NI 1 winter	-2.05	54	.045
NI 1 spring - NI 1 winter	-1.14	60	.261
QD K winter - QD K spring	-2.18	50	.034
QD K spring - QD 1 winter	-6.54	54	< .001
QD 1 winter - QD 1 spring	-2.61	60	.011
MN K winter - MN K spring	-4.64	50	< .001
MN K spring - MN 1 winter	-6.62	54	< .001
MN 1 winter - MN 1 spring	-.35	60	.726

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. K = kindergarten; 1 = first grade.

Validity

Relationships among Experimental Measures

I explored the relationships among kindergarten and first grade TEN measures with Pearson product moment correlations, and Tables 5 and 6 present these correlations. Readers will note that I presented these data in two separate tables in order to improve readability. I computed a total of 144 correlations and 108 (108/144 or 75%) were statistically significant. For a *p*-value of .05 and 200 correlations, one would expect only ten (10/200 or 5%) correlations to be due to chance. I used Cohen's (1992) criteria for effect sizes to evaluate the data. According to Cohen, for *r*, small, medium, and large effect sizes are .10, .30, and .50, respectively. Thus, there were small to large effect sizes (r = .10 - .67) among the kindergarten and first measures. I expected the strength of the relationship between the TEN measures to be medium, as the authors of the measures intended them to correspond to different early numeracy concepts.

Table 5 presents relationships among kindergarten winter 2006 TEN measures and all first grade measures. As expected, correlations among the kindergarten winter TEN measures and both winter and spring first grade TEN measures indicated medium effects and were significantly correlated for the same measure. Among all of the winter kindergarten measures, the NI and MN measures generally showed the strongest correlations with the first grade winter TEN. For both of these measures, three out of four correlations with first grade winter TEN measures showed medium to large effects. The kindergarten winter OC and NI measures each showed small to medium effect sizes for correlations with three out of four of the first grade spring measures. In contrast, the kindergarten winter QD and MN demonstrated medium to large effects with only two of

the first grade spring measures (first grade spring QD and MN). Considered overall, the kindergarten winter NI measure was most consistently related to both winter and spring first grade measures, showing significant relationships with all but one (first grade spring OC) of the other measures. When one considers these data overall, they support the hypothesis that kindergarten and first grade TEN performance would be significantly related (H02).

Table 6 presents correlations among kindergarten spring 2006 TEN measures and all first grade measures. Again, as expected, both winter and spring first grade measures generally correlated significantly with corresponding measures. For example, the strongest relationship was between the kindergarten MN spring measure and the first grade MN spring measure ($r = 0.67$, a large effect). The kindergarten spring NI measure showed nearly medium to large effect sizes for relationships with all of the first grade measures ($r = 0.29$ - 0.50). In addition, there were medium to large effects for the kindergarten spring MN measure and the first grade spring measures ($r = 0.33$ - 0.67). Further, the kindergarten spring MN measure demonstrated some of the strongest relationships with the other kindergarten and first grade measures among all kindergarten and first grade measures. Again, these data support the hypothesis (H02) that kindergarten and first grade TEN performances would be significantly related.

Table 5

Kindergarten Winter TEN and First Grade Winter and Spring TEN Correlations

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	.38**	.22	.40**	.30*	.19	.18	.49**	.39**	.28*	.33*	.32*
2		1.00	.47**	.45**	.26*	.41**	.45**	.47**	.21	.39**	.39**	.40**
3			1.00	.52**	.28*	.18	.47**	.27*	.13	.22	.42**	.35**
4				1.00	.38**	.10	.51**	.42**	.20	.15	.37**	.47**
5					1.00	.41**	.33**	.30*	.41**	.29*	.39**	.24
6						1.00	.43**	.30*	.30*	.49**	.58**	.38**
7							1.00	.43**	.11	.34**	.59**	.40**
8								1.00	.32*	.36**	.63**	.57**
9									1.00	.56**	.43**	.25*
10										1.00	.68**	.42**
11											1.00	.68**
12												1.00

Note. 1 = K OC-W; 2 = K NI-W; 3 = K QD-W; 4 = K MN-S; 5 = 1 OC-W; 6 = 1 NI-W; 7 = 1 QD-W; 8 = 1 MN-W; 9 = 1 OC-S; 10 = 1 NI-S; 11 = 1 QD-S; 12 = 1 MN-S.

K = kindergarten; 1 = first grade; OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter, S = Spring.

*$p < .05$, 2-tailed. **$p < .01$, 2-tailed.

Table 6

Kindergarten Spring TEN and First Grade Winter and Spring TEN Correlations

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	.44**	.29*	.48**	.41**	.14	.16	.30*	.40**	.31*	.26	.27*
2		1.00	.42**	.46**	.47**	.29*	.44**	.43**	.32*	.38**	.51**	.50**
3			1.00	.19	.33*	.26	.51**	.13	.04	.15	.38**	.30*
4				1.00	.22	.26	.36**	.56**	.33*	.39**	.53**	.67**
5					1.00	.41**	.33**	.30*	.41**	.29*	.39**	.24
6						1.00	.42**	.30*	.30*	.49**	.58**	.38**
7							1.00	.43**	.11	.34**	.59**	.41**
8								1.00	.32*	.36**	.63**	.57**
9									1.00	.56**	.43**	.25*
10										1.00	.68**	.42**
11											1.00	.68**
12												1.00

Note. 1 = K OC-S; 2 = K NI-S; 3 = K QD-S; 4 = K MN-S; 5 = 1 OC-W; 6 = 1 NI-W; 7 = 1 QD-W; 8 = 1 MN-W; 9 = 1 OC-S; 10 = 1 NI-S; 11 = 1 QD-S; 12 = 1 MN-S.

K = kindergarten; 1 = first grade; OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter, S = Spring.

*$p < .05$, 2-tailed. **$p < .01$, 2-tailed.

Predictive Validity of Experimental Measures

Correlation analyses. Table 7 presents correlation coefficients among the kindergarten TEN and first grade outcome measures. Overall, predictive validity correlations indicated small to large effect sizes. Out of all of the kindergarten predictor variables, MN spring performance demonstrated strong relationships with both winter (r = .52) and spring (r =. 47) first grade M-CBM performance. The OC (*range* =.31 to .38) and NI measures (*range* =.29 to .41) also evidenced medium relationships with the winter and spring M-CBM performance. I also found medium to large effect sizes for the relationship between kindergarten TEN variables and the first grade ACES-M score, with the strongest relationship occurring with the kindergarten spring MN measure (r = .58). There were medium effect sizes for the relationships between the kindergarten winter and spring QD measures (r = .35 and .34, respectively) and kindergarten spring MN measure (r = .32) and first grade end of the year overall mathematics report card grades. All other kindergarten measures correlated weakly (r < .30) with report card grades. I hypothesized (H03) that kindergarten TEN performance would be significantly related to first grade M-CBM performance. The data support this hypothesis, specifically for the OC, NI, and MN measures. Subsequent sections address hypotheses (H04-H05) for the ACES-M and report card grade outcomes.

Table 7

Relationships Among Kindergarten Experimental and First Grade Criterion Measures

	Kindergarten TEN							
	OC-W	NI-W	QD-W	MN-W	OC-S	NI-S	QD-S	MN-S
First Grade Outcomes								
M-CBM-W	.38**	.41**	.23	.42**	.31*	.37**	.13	.52**
M-CBM-S	.35**	.41**	.22	.29*	.33*	.29	.12	.47**
ACES-M	.39*	.46**	.51**	.44**	.36*	.57**	.53**	.58**
Report Card	.20	.17	.35**	.23	.04	.22	.34*	.32*

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter, S = Spring.

*$p < .05$, 2-tailed. **$p < .01$, 2-tailed

I did not include the ODRs variable in correlation analyses because only 8 participants received any ODRs. However, for descriptive purposes, Table 8 presents kindergarten TEN scores for students with and without any first grade ODRs and results of *t*-tests for each comparison. With the exception of the kindergarten spring OC measure, students with no first grade ODRs obtained higher average scores on the kindergarten TEN. Differences were most pronounced on the winter OC and spring NI measures. However, these differences were not statistically significant for any of the measures. Thus, the data do not support the hypothesis (H06) that kindergarten TEN performance would significantly predict first grade disciplinary referrals.

Table 8

Kindergarten TEN Performance of Students With and Without Office Disciplinary Referrals (ODRs) in the First Grade

Measurement Period	*Kindergarten TEN Score*		*t*	*p*
	No First Grade ODRs	One or More First Grade ODRs		
	M (*SD*)	*M* (*SD*)		
	(*n* = 53)	(*n* = 8)		
OC Winter	51.14 (17.30)	45.13 (15.82)	-.92	.36
NI Winter	48.45 (14.15)	46.18 (16.80)	-.41	.68
QD Winter	18.18 (12.97)	14.10 (11.62)	-.84	.41
MN Winter	9.10 (6.88)	7.13 (5.14)	-.76	.45
OC Spring	57.83 (16.25)	58.71 (17.34)	.13	.89
NI Spring	52.03 (13.76)	41.21 (29.87)	-1.63	.11
QD Spring	21.29 (13.21)	16.93 (9.93)	-.84	.41
MN Spring	13.29 (5.50)	10.73 (7.04)	-1.13	.27

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number.

Regression analyses. I conducted multiple regression analyses to determine which kindergarten TEN measures were the best predictors of students' mathematics computation skills, teacher ratings of mathematics skills, and end of year overall

mathematics report card grades. Prior to conducting all regression analyses, I centered all outcome variables around their respective means.

I conducted four analyses to examine the predictive validity of the experimental measures for math computation skills, as measured by M-CBM. Two analyses used data from the kindergarten TEN administered in winter to predict either winter or spring first grade M-CBM performance, and two analyses used data from kindergarten spring TEN to predict either winter or spring first grade M-CBM performance. In all four analyses, I used backwards deletion procedures to evaluate the contribution of individual predictors. Thus, I entered all of the predictors first and successively removed predictors on the basis of the lowest *B* significance levels (i.e., largest *p* value) using the criteria of $p = .10$ for inclusion in the model. Tables 9 through 12 present the results of these analyses.

Table 9 illustrates that the multiple correlation using all four winter predictor variables in the equation where winter M-CBM was the outcome variable was .51 ($p <$. 05). This result means that kindergarten winter TEN performance accounted for approximately 26% of the variance in first grade M-CBM performance. This result also suggests that other factors account for approximately 74% of the variance in M-CBM performance. When I entered all of the variables in the equation, I found that none of them demonstrated significant individual contributions. However, the backwards deletion procedures described above yielded a final model where the winter MN and NI measures were significant unique predictors of first grade winter M-CBM performance. It must be noted that the multiple correlation was lower for the final model, at .49 ($p <$. 05) than for the previous analysis. Therefore, the winter kindergarten MN and NI measures account for approximately 24% of the variance in first grade winter M-CBM performance.

Table 10 illustrates the percentage of variance in spring M-CBM accounted for by the kindergarten winter TEN. Overall, all kindergarten winter TEN performance scores accounted for approximately 21% of the variance in first grade spring M-CBM performance. Following the removal of the kindergarten winter QD and MN measures, according to backwards deletion procedures described above, I found a final model with kindergarten winter NI and OC predictors. Only NI contributed uniquely and significantly to first grade spring M-CBM performance. These finding are consistent with the correlation data in Table 5 where the kindergarten winter NI measure correlated more highly with first grade spring M-CBM performance than any of the other kindergarten TEN measures.

Tables 11 and 12 demonstrate that the kindergarten spring TEN predictors together accounted for 29% of the variance in first grade winter M-CBM performance and 23% of the variance in first grade spring M-CBM performance. The only measure that uniquely and significantly contributed to both winter and spring first grade M-CBM performance was the kindergarten spring MN measure. I removed all other measures from the model, according to the aforementioned decision rules. Removal of these measures produced minimal difference in the percentage of variance accounted for, changing variance accounted for from 29% to 27% for winter M-CBM and from 23% to 22% for spring M-CBM.

These results also support the hypothesis (H03) that kindergarten TEN performance would be significantly related to first grade M-CBM. In particular, the kindergarten winter MN and NI measures were significant predictors of first grade winter M-CBM, while the kindergarten winter NI measure was a significant predictor of spring

first grade M-CBM, and the kindergarten spring MN measure was a significant predictor of both winter and spring first grade M-CBM.

Table 9

Summary of Multiple Regression Analysis for Kindergarten Winter TEN Variables Predicting First Grade Winter M-CBM

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-W	.027	.022	.165			1.22	.23
NI-W	.049	.028	.252			1.74	.09
QD-W	-.015	.033	-.068			-.46	.65
MN-W	.117	.064	.278	.51*	.26	1.84	.07
OC-W	.028	.022	.169			1.27	.21
NI-W	.045	.027	.231			1.70	.10
MN-W	.106	.058	.251	.51*	.26	1.82	.07
NI-W	.054	.026	.274*			2.06	.04
MN-W	.126	.056	.299*	.49*	.24	2.25	.03

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter.

*p < .05

Table 10

Summary of Multiple Regression Analysis for Kindergarten Winter TEN Variables Predicting First Grade Spring M-CBM

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-W	.038	.026	.208			1.50	.14
NI-W	.065	.033	.295			1.97	.06
QD-W	-.001	.038	-.006			-.04	.97
MN-W	.036	.073	.075	.46*	.21	.48	.63
OC-W	.038	.025	.208			1.52	.14
NI-W	.064	.031	.293*			2.08	.04
MN-W	.034	.067	.073	.46*	.21	.51	.61
OC-W	.042	.024	.227			1.74	.09
NI-W	.070	.029	.318*	.46*	.21	2.43	.02

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter.

*p < .05

Table 11

Summary of Multiple Regression Analysis for Kindergarten Spring TEN Variables Predicting First Grade Winter M-CBM

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-S	.006	.025	.033			.23	.82
NI-S	.026	.024	.155			1.09	.28
QD-S	.001	.028	.003			.03	.98
MN-S	.215	.070	.435*	.54*	.29	3.05	< .01
OC-S	.006	.024	.033			.24	.81
NI-S	.026	.023	.155			1.13	.26
MN-S	.215	.070	.435*	.54*	.29	3.08	< .01
NI-S	.028	.022	.165			1.26	.21
MN-S	.221	.065	.446*	.54*	.29	3.41	< .01
MN-S	.258	.058	.521*	.52*	.27	4.44	<.01

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. S= Spring.

*p < .05

Table 12

Summary of Multiple Regression Analysis for Kindergarten Spring TEN Variables Predicting First Grade Spring M-CBM

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-S	.025	.025	.128			.86	.39
NI-S	.003	.028	.018			.12	.90
QD-S	.005	.032	.019			.15	.88
MN-S	.217	.082	.393*	.48*	.23	2.65	.01
OC-S	.026	.028	.133			.93	.34
QD-S	.006	.031	.022			.18	.86
MN-S	.220	.077	.398*	.48*	.23	2.85	.01
OC-S	.027	.027	.138			1.00	.32
MN-S	.221	.076	.400*	.48*	.23	2.89	.01
MN-S	.257	.067	.466*	.47*	.22	3.83	<.01

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. S = Spring.

*p < .05

I also conducted multiple regression analyses to determine which kindergarten TEN measures were the best predictors of teacher ratings of students' mathematics skills on the ACES-M. I conducted two analyses: one with kindergarten winter predictors and one with kindergarten spring predictors. Again, I entered all of the predictors first in order to evaluate the overall contribution of the measures, and then successively removed predictors on the basis of *B* significance levels, using a criterion of $p = .10$ for inclusion, in order to identify the best individual predictors. Tables 13 and 14 summarize the results of these analyses. The multiple correlation values for these equations were .62 ($p < .05$) using the winter predictors and .77 ($p < .05$) using the spring predictors. Thus, in the final model, kindergarten winter and spring TEN performance accounted for 35% and 53% of the variance on teacher ratings on the ACES-M, respectively. When I examined kindergarten winter predictors, the OC and QD measures made significant individual contributions to ratings on the ACES-M. For the kindergarten spring predictors, the QD, and MN measures contributed uniquely and significantly to the ACES-M ratings.

Thus, the data also supported the hypothesis (H04) that kindergarten TEN would significantly predict first grade teacher ratings of mathematics skills. Specifically, the kindergarten winter OC and QD measures and the kindergarten spring QD and MN measures made unique significant contributions to first grade ACES-M ratings.

Table 13

Summary of Multiple Regression Analysis for Kindergarten Winter TEN Variables Predicting First Grade ACES-M

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-W	.070	.054	.211			1.30	.20
NI-W	.068	.069	.173			.99	.33
QD-W	.146	.079	.328			1.85	.08
MN-W	.093	.155	.109	.62*	.38	.60	.55
OC-W	.079	.051	.238			1.55	.13
NI-W	.076	.067	.269			1.13	.27
QD-W	.165	.072	.370*	.61*	.37	2.29	.03
OC-W	.098	.049	.294*			2.01	.05
QD-W	.200	.065	.449*	.59*	.35	3.06	< .01

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W= Winter.

*p < .05

Table 14

Summary of Multiple Regression Analysis for Kindergarten Spring TEN Variables Predicting First Grade ACES-M

Variables in the Equation	*B*	*SE B*	*β*	*Multiple R*	R^2	*t*	*p*
OC-S	-.026	.051	-.074			-.51	.61
NI-S	.100	.049	.294			2.03	.05
QD-S	.175	.057	.395*			3.08	.01
MN-S	.414	.145	.415*	.77*	.59	2.87	.01
NI-S	.094	.047	.276			1.99	.06
QD-S	.171	.055	.385*			3.08	.01
MN-S	.389	.134	.390*	.77*	.59	2.91	.01
QD-S	.198	.056	.447*			3.52	< .01
MN-S	.504	.127	.505*	.73*	.53	3.98	< .01

Note. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. S = Spring.

*p < .05.

I used ordinal regression analyses to determine which kindergarten TEN measures were the best predictors of students' end of the year overall mathematics report card grade. One uses this approach when the dependent variable is an ordinal variable, or a variable that uses rank orderings where the difference between each category is unknown (Norusis, 2008). Further, in ordinal regression, it is of interest to determine the probability of obtaining a certain score or lower. Teachers recorded report card grades as a 2 (*meets grade level expectations with assistance*), 3 (*meets grade level expectations*), or 4 (*above grade level expectations*). Ordinal regression analyses examined the probability that the experimental variables could predict the probability of obtaining a grade of 2, or a grade of 3 or higher. Examining the data this way was particularly appropriate considering that, as illustrated in figures 2 and 3, students who received grades of 3 *(as meets grade level expectations*) and 4 (*above grade level expectations*) performed similarly on the experimental measures. In other words, there appeared to be two distinct outcomes: (a) receiving a grade of 2, and (b) receiving a grade or 3 or 4.

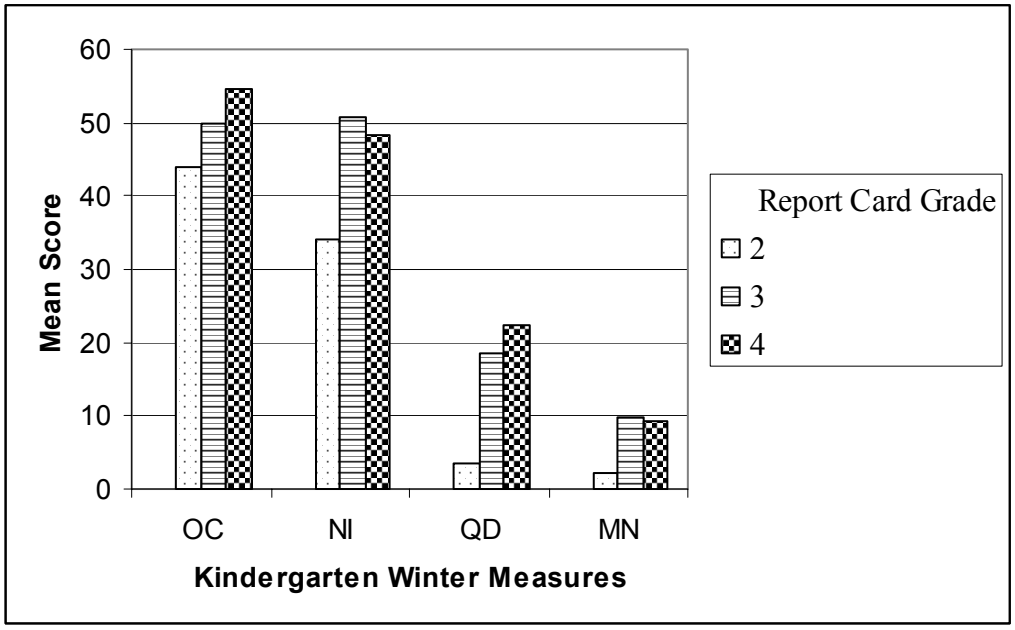


Figure 2. Winter Kindergarten TEN Performance by Report Card Grade

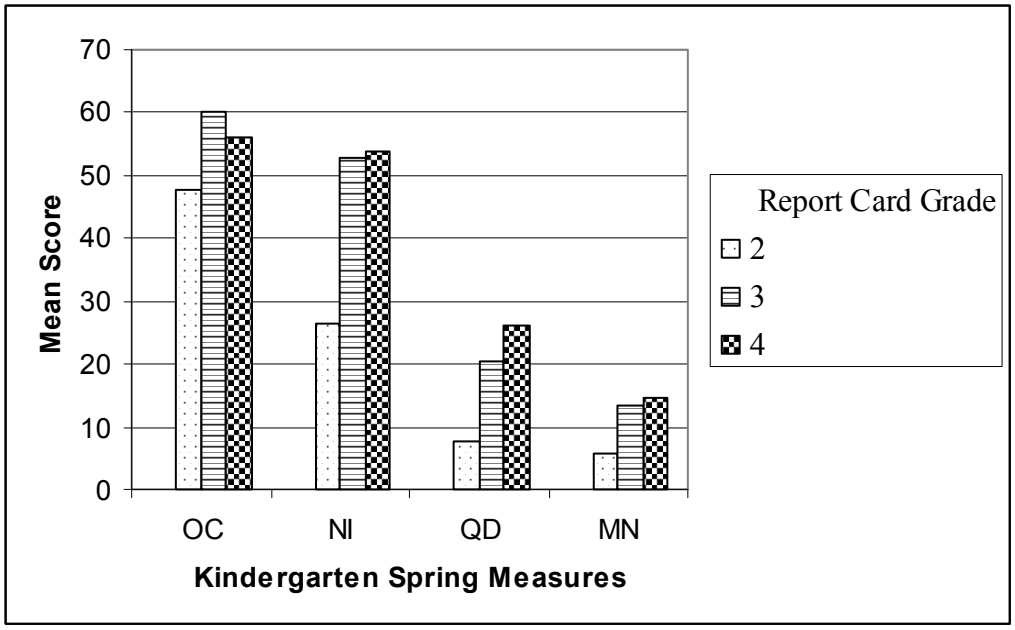


Figure 3. Spring Kindergarten TEN Performance by Report Card Grade

First, I entered all of the winter kindergarten TEN measures as predictor variables. The overall model was not was not significant, χ^2 (4, N = 61) = 8.91, p > .05. However, Table 15 illustrates that in the examination of the individual predictors, the QD measure demonstrated significant unique contributions to the outcome. Backwards elimination procedures removed non-significant predictors successively from the equation. A final model that included only the kindergarten winter QD measure was significant, χ^2 (1, N = 61) = 8.33, p < .05. In this model, kindergarten winter QD performance accounted for 16% of the variance in first grade end of year overall mathematics report card grade.

In the next analysis, I entered kindergarten spring TEN measures into the equation. The overall model was statistically significant, χ^2 (4, N = 61) = 16.90, p < .05, indicating that kindergarten spring TEN performance collectively predicted first grade end of the year overall mathematics report card grade. Specifically, kindergarten spring performance on the TEN measures accounted for 33% of the variance in first grade report card grade. Two of the measures, QD and MN made significant individual contributions to the outcome. Again, I used backwards deletion procedures to determine whether any other measures would emerge as significant predictors when I removed those measures with the highest p-values. Still, only the QD and MN measures remained as significant individual predictors, accounting for 24% of the variance in report card grade. These results mirror the results of correlational statistics in Tables 5 and 6 where the winter QD and spring QD and MN measures were most highly related to report card grades. Although only the QD and MN measures made unique contributions to this outcome; it may be more beneficial to consider the model utilizing all four kindergarten spring TEN predictors, which accounted for an additional 9% of the variance in report card grades.

Overall, the results support the hypothesis (H05) that kindergarten TEN performance would significantly predict first grade end of year overall mathematics report card grades. In particular, this hypothesis is supported for the kindergarten winter QD measure and the kindergarten spring QD and MN measures.

Table 15

Summary of Ordinal Regression Analysis for Kindergarten Winter TEN Variables Predicting Report Card Grade

Variables in the Equation	*B*	*SE B*	*OR*	R^2_N	*df*	*p*
OC-W	.013	.019	.987		1	.48
NI-W	.000	.024	1.000		1	.99
QD-W	.064*	.029	.938		1	.03
MN-W	-.004	.053	1.00	.174	1	.94
OC-W	.013	.018	.987		1	.47
QD-W	.064*	.028	.938		1	.02
MN-W	-.002	.052	1.00	.174	1	.93
OC-W	.012	.017	.988		1	.46
QD-W	.063*	.025	.939	.174	1	.01
QD-W	.066*	.024	.936	.163	1	.01

Note. OR (Odds Ratio) = e^{-B}. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. W = Winter.

*p< .05.

Table 16

Summary of Ordinal Regression Analysis for Kindergarten Spring TEN Variables Predicting Report Card Grade

Variables in the Equation	*B*	*SE B*	*OR*	R^2_N	*df*	*p*
OC-S	-.045	.024	1.046		1	.06
NI-S	.033	.022	.967		1	.13
QD-S	.063*	.028	.939		1	.02
MN-S	.141*	.067	.868	.325	1	.04
OC-S	-.036	.022	1.037		1	.11
QD-S	.069*	.027	.933		1	.01
MN-S	.167*	.066	.846	.284	1	.01
QD-S	.058*	.026	.944		1	.02
MN-S	.117*	.055	.890	.235	1	.03

Note. OR (Odds Ratio) = e^{-B}. OC= oral counting; NI = number identification; QD = quantity discrimination; MN = missing number. S = Spring.

*p< .05.

Visual Quantity Discrimination

I created the Visual Quantity Discrimination (VQD) measure for this study in order to explore students' ability to correctly discriminate quantities of object sets presented symbolically, as opposed to written numbers. Table 17 shows small to large effects for relationships between first grade VQD scores with scores from the first grade TEN measures and first grade outcomes. I computed a total of 28 correlations, 18 of which were statistically significant. Given a *p*-value of .05, one would expect only one out of 20 (1/20) correlations to be significant by chance. Among the first grade TEN measures, the VQD measure showed the strongest relationship with the first grade QD (r = .46 for winter measures; r = .56 for spring measures). One would expect this finding since the QD and VQD measures examine a closely related construct; although with different types of procedures (i.e., symbolic *vs.* written).

There were small to large effect sizes for the first grade winter VQD measure and first grade outcomes, and correlations ranged from .18 to .54. For the first grade winter VQD measure, the weakest relationship was with winter M-CBM (r = .18) and the strongest relationship was with report card grades (r = .54). Compared to the first grade winter VQD measure, correlations among the first grade spring VQD measure and first grade outcomes were lower and ranged from .19 to .28. The first grade spring VQD measure showed the weakest relationship with report card grades (r = .19) and the strongest with M-CBM (r = .28), although both of these relationships are relatively weak.

Table 17

Visual Quantity Discrimination Correlations

	VQD-Winter	VQD-Spring
OC-Winter	.31*	.33**
NI-Winter	.12	.35**
QD-Winter	.46**	.28*
MN-Winter	.16	.26*
VQD-Winter	1.00	.44**
OC-Spring	.19	.28*
NI-Spring	.19	.37**
QD-Spring	.43**	.56**
MN-Spring	.21	.28*
VQD-Spring	.44**	1.00
M-CBM-Winter	.18	.28*
M-CBM-Spring	.37*	.28*
ACES-M	.31*	.23
Report Card Grade	.54*	.19

*p < .05. **p < .01.

Summary

This dissertation generated seven hypotheses and found support for six out of the seven hypotheses tested. The first hypothesis was that students would make significant progress on the TEN measures from kindergarten to first grade. The results showed that for each measure, growth was not due to chance and thus, provided support for this hypothesis. Next, I hypothesized that kindergarten TEN performance would be significantly related to first grade TEN performance. Correlational data also provided support for this hypothesis, as there were small to large effect sizes for relationships between kindergarten and first grade measures. The next three hypotheses stated that kindergarten TEN would significantly predict first grade mathematics computation skills, as measured by M-CBM; first grade teacher ratings of mathematics skills; and end of the year overall mathematics report card grades. Results of regression analyses provided support for all three of these hypotheses.

The sixth hypothesis was that kindergarten TEN performance would significantly predict discipline referrals in the first grade. This hypothesis was not supported. Although there seemed to be some qualitative differences in the scores of students who received no discipline referrals in the first grade compared to students who received one or more discipline referrals; these differences were not statistically significant. This finding may have been due to the fact that only eight students in the sample received discipline referrals.

The last hypothesis stated that the first grade VQD measure would be related to first grade TEN performance and first grade outcomes. Data partially supported this

hypothesis as there were small to large effect sizes for relationships between the VQD measure and the first grade TEN measures and other first grade outcomes.

CHAPTER V

Discussion

While school psychologists have long studied CBM, researchers have only recently begun to examine the utility of this technique for assessing early mathematics skills (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Daly et al., 1997; Floyd, et al., 2006; VanDerHeyden et al., 2001; VanDerHeyden et al., 2004). Specifically, there have been studies on CBM of early numeracy skills that explored the predictability of early numeracy assessments with standardized assessments (e.g., Chard et. al., 2005; Clarke & Shinn, 2004), tested the sensitivity and specificity of early math screeners for special education services (e.g., VanDerHeyden et. al., 2003), examined the reliability of these measures (Chard et al, 2005; Clarke & Shinn, 2004; Petreshock et al., 2006), and correlated performance with computation M-CBM assessments (e.g., Daly et al., 1997).

Findings from this preliminary research suggest that TEN measures provide a range of test-retest (r = .46 to .86) and alternate form (r = .71 to .99) reliabilities and indicate significant predictive validity relationships with M-CBM, The Woodcock Johnson-III Applied Problems subtest (Woodcock, McGrew, & Mather, 2001), and the Number Knowledge Test (Case & Okamoto, 1996) (Chard et al., 2005; Clarke & Shinn, 2004; Petreshock et al., 2006). Similar measures of number reading and counting numbers (Daly et al., 1997) and counting skills, circling numbers, writing numbers, and drawing objects given a specific number (VanDerHeyden et al., 2001) demonstrate comparable reliability and validity evidence. Sensitivity and specificity of preschool measures involving choosing numbers, naming numbers, counting objects, and object discrimination suggest that these measures fulfill the requirements of CBM tools in that

they exhibited small changes in growth over time and were able to identify students in need of intervention in kindergarten (VanDerHeyden et al., 2004). Such research is critical given U.S. students' current performance in mathematics and educational movements calling for increased accountability and early identification and remediation of academic skills difficulties.

The purpose of this dissertation was to support and extend existing findings on the technical features of CBM of mathematics in the early school years by examining the predictive validity of the TEN measures (Clarke & Shinn, 2004), which assess early numeracy skills, longitudinally. Previous research did not examine student performance on these measures longitudinally. Longitudinal research is important because, although legislators and school personnel consider formative evaluation, or assessment incorporating frequent progress monitoring, to be a crucial component of early intervention; there have been only two studies that examined student growth on TEN over time (Chard et al., 2005; Clarke & Shinn, 2004). One should also note, even more specifically, that no studies have examined numeracy growth from kindergarten to first grade.

Previous research has also not addressed the ability of kindergarten TEN performance to predict first grade TEN performance or other broader first grade outcomes, in particular, contextually-relevant outcomes. According to VanDerHeyden et al. (2006) a goal of CBM is to collect ecologically-valid assessment data, but current research on early numeracy measures has primarily focused on relationships with standardized test data. Finally, it was important to explore which TEN measures served as the best predictors of first grade outcomes given that Clarke and Shinn (2004) suggest

that research has not yet determined whether one can assess early numeracy with a single indicator or if one must use multiple measures.

In order to begin to address these gaps in the literature, one goal of the current dissertation was to examine whether kindergarten TEN performance was related to first grade TEN performance. Secondly, the dissertation explored student progress on the TEN measures from kindergarten to first grade. Another goal of the current study was to identify which kindergarten measures were the best predictors of first grade computation skills, teacher ratings, end of year overall mathematics report card grades, and discipline referrals. Finally, an additional goal of the current study was to examine the technical properties of a measure of visual quantity discrimination.

I hypothesized that students would demonstrate significant progress over time and that kindergarten and first grade TEN performance would be significantly related and show medium to large effect sizes. There were several hypotheses regarding the predictive validity of the measures for first grade outcomes. The first hypothesis was that that performance on the kindergarten TEN measures would predict first grade computations skills, as measured by M-CBM. I also believed that kindergarten TEN performance would significantly predict first grade teacher ratings of students' math skills on the ACES-M, students' first grade end of year overall math report card grade, and the number of discipline referrals students received. Due to the exploratory nature of the study, I used several analyses, employing correlation and regression techniques, to determine which particular kindergarten measures predicted the various first grade outcomes. The results of this study confirmed these hypotheses for some, but not all, of the TEN kindergarten measures. Further, different kindergarten measures predicted

different first grade outcomes.

Sensitivity of TEN Measures

In support of the hypothesis that students' performance would increase significantly from kindergarten to first grade, students made significant gains on all the experimental measures. It was important to explore the ability of TEN to measure student growth over time, since the authors of these measures designed them to monitor progress in a formative evaluation framework. Students improved on all TEN measures over a 68-week period extending from the winter of 2006 to the spring of 2007, with growth not due to chance. Specifically, students showed the greatest improvement on the OC (0.43 digits per week) measure, followed by the QD (0.24 digits per week), MN (0.15 digits per week), and NI (0.12 digits per week) measures, respectively. These findings are partially consistent with previous research. For example, in Chard et al.'s (2005) research, kindergarten students made the most growth on the NI, followed by the MN, and QD measures. Chard et al. observed the same pattern in their sample of first graders.

As Chard et al. (2005) did not include a counting measure in their analysis of student progress; I cannot compare my findings for this measure with theirs. However, the Clarke and Shinn's (2004) study did include the OC measure, and consequently, found results similar to the present study in that the first graders showed the greatest growth on the OC measure. Following the OC measure, Clarke and Shinn found that first grade growth was greatest on the NI, QD, and MN measures over a 13-week time period. Taken together, the most obvious inconsistency between the findings from the present study and those of Clarke and Shinn and Chard et al. was that growth on the NI measure was greater than growth on the QD and MN measures in the latter two studies, whereas

students showed the least progress on the NI in the current study. Still, the amount of student growth per week, in terms of digits gained, roughly matched student gains recorded in previous research and was relatively slow for all measures (Clarke & Shinn; Chard et al.). Specifically, results from all three studies indicate that it would take at least two and up to ten weeks, depending on the TEN measure, for students to increase their scores by 1 digit. This is an important consideration for educators to keep in mind when examining student progress on these measures. Nevertheless, TEN growth rates are very consistent with research on rates of progress with M-CBM that shows that weekly growth on M-CBM ranges from approximately .2 to .7 digits per week across grades 2 through 6 (Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993).

When I examined comparisons of performance between testing time pairs of the current study, I found that students' growth on the OC and MN measures was significant within kindergarten and but not within first grade. In contrast, growth on QD was significant within first grade but not within kindergarten. Student growth on the NI measure was not significant for any of the testing time pairs. Thus, while all of the measures were able to detect significant growth over the course of the study; the OC and MN measures were sensitive within kindergarten and the QD measure was sensitive within first grade level. Again, these results correspond with those of Chard et al. (2005), who also found that the kindergarten MN measures was sensitive to growth; and along with Clarke and Shinn (2004), found that the first graders made significant progress on the QD measure. These findings suggest developmental differences in skill development or curricular differences over time. In particular, it appears that measures of counting skills (i.e., OC and MN) are appropriate in kindergarten, but possibly result in a ceiling

effect in the first grade. In contrast, there may be a floor effect for the kindergarten QD measure, which appears to be better matched to skills acquired at the first grade level.

Predicting First Grade TEN from Kindergarten TEN Performance

I expected that kindergarten and first grade performance on TEN would be correlated, because the primary difference between these measures is the upper numerical limit. That is, I expected students to perform all number sense tasks with numbers up to 20 for first grade and numbers up to 10 for kindergarten. The study results showed small to large effect sizes among kindergarten and first grade measures, indicating that these measures collectively assess a broader construct of early mathematics skills. Relationships among corresponding kindergarten and first grade measures (e.g., kindergarten OC and first grade OC) were often somewhat stronger than the relationships among the different measures overall (i.e., OC, NI, QD, and MN). This is not an unexpected result given that Clarke and Shinn (2002) designed the combined set of measures to measure the broader construct of number sense, but they designed each individual measure to assess a specific number sense concept.

The small to large effect sizes (r = .10 to .67) among kindergarten and first grade TEN measures in this study were only partially consistent with results of previous research on these measures (Clark & Shinn, 2004; Petreshock et al., 2006). VanDerHeyden, Broussard, and Cooley (2007) examined performance on similar measures investigated longitudinally from pre-school to kindergarten and found correlations comparable to my findings (r = .31 to .66). However, my findings are inconsistent with Clarke and Shinn who found large to very large effect sizes among first grade TEN probes (r =.55 - .93). The disparity in my findings and Clarke and Shinn's

findings is not surprising given that I expected only medium effect sizes for kindergarten and first grade measures.

It is noteworthy that my findings more closely matched those from my pilot study with kindergarten students, where correlations among kindergarten TEN measures ranged from .23 to .61. The same students participated in the pilot study and the dissertation; thus, the results indicate that these students' performances demonstrated consistent relationships among the TEN measures. In addition, differences between the results of this study and those of Clarke and Shinn's (2004) work might be due to the fact that this study utilized data collected over different grade levels. It is possible that both the longer time span of the current study as well as the fact that the kindergarten and first grade measures contained slightly different content (numbers up to 10 or up to 20, respectively) may have resulted in lower correlations among kindergarten and first grade measures. Overall though, it appears that the TEN measures are useful in predicting number sense skills from kindergarten to first grade.

It is interesting to note that the kindergarten MN measure showed some of the strongest relationships with the first grade TEN measures and thus, may tap into several skills associated with number sense. This finding paralleled my pilot research results where kindergarten MN demonstrated the strongest relationship with the other kindergarten TEN measures. Therefore, for kindergarten students, the ability to identify the missing number in a string of numbers appears to be one of the best indicators of both current and future overall number sense.

However, for first graders, QD demonstrated the strongest relationships with the other measures. Clark and Shinn (2004) also found this when they studied concurrent

validity correlations of TEN measures with first grade. It seems reasonable to conclude that distinct grade level differences may exist with respect to which skills are the best indicators of number sense. Specifically, the ability to identify the missing number in a string of numbers appears to be the most representative skill at the kindergarten level with the ability to discern which of two numbers is larger being the better indicator at the first grade level. These data match the aforementioned findings regarding the sensitivity of TEN where the MN measure was more sensitive to growth at the kindergarten level and the QD measure was more sensitive to growth at the first grade level. This formulation corresponds with Gersten et al.'s (2007) assertion that magnitude comparison is a recurrent skill in operational definitions of number sense and that developing mental number lines is essential for mathematical proficiency. Therefore, the skills measured by the MN and QD measures, in particular, are well matched to theory on number sense.

Predicting First Grade Outcomes

Predicting computation. The relationship between kindergarten performance on TEN and first grade math computation, as measured by CBM-M, ranged from weak to strong depending on the early numeracy skill examined. Specifically, the kindergarten MN measure generally demonstrated the strongest relationship with first grade computation skills; although the OC and NI measures also showed medium effect sizes. Moreover, TEN performance accounted for a significant (R^2 = 0.21 - 0.27, a medium to large effect size) amount of variance in first grade M-CBM performance. This finding is not necessarily surprising given that Chard et al. (2005) found that TEN accounted for 44% of variance in the Number Knowledge Test (Okamoto & Case, 1996), which measures a variety of number sense skills. Consistent with the correlation analyses,

kindergarten MN was a significant and unique predictor in three out of the four analyses that examined the contribution of kindergarten TEN to first grade M-CBM. Furthermore, the kindergarten spring MN measure was the only significant individual contributor to first grade M-CBM performance.

Taken together with the finding that MN also yielded the strongest correspondence across years (from kindergarten to first grade) and was sensitive to growth at the kindergarten level, one could consider that MN has strong support as an early indicator. If kindergarten MN is indeed the best overall indicator of early mathematics skills, it is not surprising that it would be the best indicator of skills that build upon pre-requisites, namely computation skills. In addition, based on the fact that students received the lowest scores on the MN measure, one can consider MN to be the most difficult task among the measures. In this way, scores on MN may be indicative of whether students can complete higher level skills such as computation. Similarly, from a cognitive perspective, one would expect students who have greater automaticity with basic skills such as number line skills, to be more proficient with higher-level skills, such as computation, based on theory stating that automaticity or fluency with basic facts makes other cognitive resources (e.g., attention, working memory) more available for more complex tasks (Poncy, Skinner, & Jaspers, 2006).

Findings regarding the MN measure and computation skills are not consistent with previous research by Clarke and Shinn (2004) who found that the QD measure demonstrated the best predictive validity for M-CBM performance for first graders over one school year. Since Clarke and Shinn's research involved first grade students, it is not surprising that the QD measure was a better predictor of M-CBM, as the QD measure

also appears to be the best overall indicator of early mathematics for first grade students according to the present study. Another possible reason that QD measured at the kindergarten level did not relate as well to mathematics skills as did QD measured at the first grade level is that this study did not include more global measures of mathematics as outcome variables, and QD may serve as a better predictor of these skills. That is, mathematics minimally consists of computation and applications (e.g., word problems, algebra, data analysis, measurement, geometry, and patterns) skills and although the constructs representing computation and applications skills are highly related, they contribute separately to mathematics knowledge (Thurber, Shinn, & Smolkowski, 2002). Indeed, in my pilot study, the QD measure was more strongly related to performance on an applied problems criterion measure than the MN measure.

In summary, it seems plausible that the MN measure may serve as stronger indicator of computations skills when assessment occurs in kindergarten while the QD measures may be a better indicator at the first grade level, at least according to Clarke & Shinn (2004). Thus, an important contribution of the present study is the finding that the best indicator of mathematics computation performance for kindergarten students is not the same as the best predictor of mathematics computation performance for first grade students. Interestingly, Gersten et al. (2005) describe the MN measure as measure of counting knowledge and thus, it is not surprising that this type of measure might be a better indicator for students who are presumably in the earlier stages of number sense. It seems that counting and number line knowledge would be important pre-requisite skills for addition and subtraction that involve counting up or down from some number or understanding how many digits to move up or down a number line. Indeed, Gersten and

colleagues (Gertsten & Chard, 1999; Gersten et al., 2005) have described cognitive abilities related to symbolically or visually storing numerical information, such as mental number lines, as critical for both conceptual understanding and problem-solving in mathematics.

Predicting teacher ratings of mathematics performance. I included teacher ratings on the ACES-M in order to obtain a standardized measure of teachers' perceptions of students' mathematics skills. All of the kindergarten TEN measures showed medium to large effect sizes with teacher ratings on the ACES-M. It is interesting to note that while teacher ratings and report card grades were significantly related; the TEN measures were more highly correlated with the ACES-M than with report card grades. This may be due to the fact that the standardized ACES-M ratings are more reliable than report card grades. In addition, report card grades may have accounted for a wider range of skills and attributes than the ACES-M, which may not have been as well matched to those skills represented in TEN.

Kindergarten winter and spring TEN performance collectively accounted for a considerable amount of the variance in ACES-M ratings. Specifically, the kindergarten winter and spring TEN measures accounted for 35% and 53% of the variance in first grade ACES-M ratings, respectively. This finding also supports the notion that, all together, TEN measures the broader construct of mathematics skills, at least according to teacher reports. Regression analyses indicated that when I entered the winter measures into the equation, the OC and QD measures were significant predictors of ACES-M scores, while the spring QD and MN measures emerged as significant predictors of ACES-M scores. Thus, the QD shows the most consistency over time as a significant

predictor of teacher ratings on the mathematics portion of the ACES.

In addition, other early numeracy skills, in particular those related to counting - OC and MN - contributed to teacher ratings depending on the assessment point. Oral counting skills appear more important earlier in kindergarten and then, MN performance becomes more important as students reach the end of the year. However, these measures appear to tap a similar skill based on the significant relationship I found between OC and MN measures. The fact that the OC and MN measures are related is also supported by Gersten et al.'s (2005) assertion that the MN measure assesses counting knowledge. Support for the OC and MN measures as early indicators of mathematics skills is also consistent with the results of research suggesting that counting skills are fundamental to developing other types of mathematical understanding (Dowker, 2005; Gersten et al., 2007) and the work of Floyd and colleagues (2006), who found that oral counting fluency may be the best indicator of early numeracy in the preschool years (Floyd et al., 2006). Gersten et al. (2007) also specify that sequence counting, which refers to reciting the number names, is an important skill for students to master in preschool and later, counting strategies (e.g., counting up from a given number) become more important for mathematical skill development.

Predicting report card grades. With respect to report card grades, the winter and spring QD measure performance demonstrated significant correlations and medium effect sizes with the end of year overall mathematics report card grade (i.e., 1 [*does not meet grade level expectations*] to 4 [*above grade level expectations*]) along with the spring MN measure. When submitted to ordinal regression analyses, the winter and spring TEN collectively accounted for a notable 16% and 33% of the variance in end of the year

report card grade for overall mathematics, respectively. Evaluation of the spring predictors showed that the QD and MN measures were both significant individual predictors. It is interesting to note that these results also correspond to findings for the ACES-M. That is, in this study, the MN and QD measures were both useful single indicators of how students perform within the school context based on teacher ratings and report card grades.

Visual Quantity Discrimination

A final aspect of the study requiring discussion is the inclusion of the VQD measure, which I constructed for the purposes of this study in order to assess the importance of students' ability to discern which of two sets of objects was larger than the other as an indicator of number sense. As I created the VQD measure after the pilot study, I was only able to administer it when students were in first grade. Therefore, I could not include VQD as predictor variable in analyses testing which kindergarten measures contributed to first grade outcomes. Nevertheless, data from the first grade administrations indicates that this measure failed to show adequate test-retest reliability (r = .44) according to criteria set by Salvia and Ysseldyke (1998), which state that reliability estimates of .60 or higher are required for making decisions about group of students.

As such, one would not typically conduct validity analyses. However, it is worth nothing that the VQD measure did show some significant relationships, and small to large effect sizes, with kindergarten and first grade measures and with first grade outcomes. I expected higher correlations given that research identifies the ability to use multiple representations of the same number and compare quantities as accepted components of number sense (Gersten et al., 2005) and since the skill the VQD assesses is similar to that

assessed by the QD measure, with a different format. Further, the fact that one would expect the VQD measure to show stronger relationships with the other TEN measures is especially true given that Clarke and Shinn (2004) and Chard et al. (2005) in part established the construct validity of the TEN by exploring the relationships among TEN performance and performance on The Number Knowledge Test, which requires students to engage in tasks involving comparing sets of objects. In addition, Floyd et al. (2006) found data to support the technical properties of a similar measure where students made judgments about the magnitude of sets of circles.

The most likely explanation for my findings for the VQD measure is that there should be more careful construction of a measure to measure students' ability to compare object sets of differing quantities. Indeed, anecdotal reports from examiners indicated that the VQD measure was the most difficult to administer and score due to the fact that one had to simultaneously observe students' rapid pointing and record these responses on the corresponding examiner sheet. Floyd et al.'s (2006) measure contained only one item per page, which may have eased administration and yielded more reliable performance.

Limitations

As with any research, this study had several limitations. Since this study employed a longitudinal design, several students left the participating school district from kindergarten to first grade, resulting in an overall smaller sample size. Although the final sample size met requirements for statistical power and the attrition group did not differ significantly from the final group on the majority of important demographic variables; the benefits of having a larger sample were lost. In relation, the relatively low return rate of ACES-M forms resulted in a smaller than desired, albeit adequate, sample size for

analyses. Given these limitations, readers should only generalize the results of the study to populations similar to the sample with respect to demographic characteristics such as age, race/ethnicity, and geographic location. While this study used a diverse group of students and appears more representative of the U.S. population at large than previous studies; practical constraints limited the sample in this respect as well. Most notably, participants came from only a single district in one region.

Another limitation of this study relates to the time-frame over which I collected data. While this study provides preliminary evidence for predictive validity of TEN over two academic years, the ability of kindergarten TEN to predict academic and behavioral outcomes later in schooling is unknown. Also related, the study used only two assessment points during each school year. Typically, one conducts benchmark assessments three times yearly (Batsche et al., 2006; Clarke & Shinn, 2002) in order to gain periodic data on the progress of all students. The inclusion of fall data would have provided a more complete picture of growth from the beginning of the school year to the end. The setting in which I conducted the study, in school hallways, may have also limited the findings. As described in Chapter III, examiners made attempts to minimize disruptions during assessment times. In addition, CBM is typically collected in informal settings such as classrooms (Derr-Minneci & Shapiro, 1992). However, it was impossible to control for all hallway activity, which may have negatively impacted some students' performances.

Last, but not least, additional types of outcome measures, both academic and behavioral, would have strengthened the research and should be strongly considered in future studies. In particular, ODRs proved to be a poor measure of problematic behavior, because so few students received any at all. With respect to academic measures, the

current study measured only mathematics computation skills. Inclusion of a global measure of mathematics concepts or a separate measure targeting applications skills would have enhanced the conclusions regarding TEN's predictions of mathematics performance.

Future Research

Given the aforementioned limitations of this study, there are several implications for future research. First, replication of the current research would provide additional support for TEN. However, future studies will be strengthened by employing a larger participant group. In addition, to provide greater generalization future studies should have a more diverse group of participants.

Second, the inclusion of additional assessment points might provide useful information about student growth on the TEN measures as they are designed to be used for formative evaluation. Measuring performance at the commonly recommended benchmark points might provide greater understanding of expected rates and patterns of growth and therefore, help solidify guidelines for making decisions about progress. Similarly, no one has explored the use of TEN for frequent progress monitoring in the context of intervention. Future studies might examine the weekly TEN performance of students receiving early numeracy intervention. These types of data would also provide guidelines for decision making specifically related to progress in response to more intensive intervention. Third, a more controlled testing setting may be beneficial to future studies.

Fourth, future research might examine whether and how kindergarten TEN performance predicts outcomes later in schooling, particularly student performance on

high stakes testing that occurs during middle elementary school years or special education eligibility. For example, a study might involve collecting follow-up student record data such as state assessment results, standardized mathematics test results, or other data related to placement decisions such as report card grades, classroom exam grades, or special education referrals. Correlation or regression analyses might then be used to explore the relationship between kindergarten or first grade performance and later outcomes. These findings would provide greater evidence for the utility of TEN in early intervention efforts. These types of data might also help determine whether the TEN measures are useful for identifying children with specific learning disabilities. That is, this study demonstrates that TEN is an indicator of later mathematics performance, and there is some evidence that early numeracy CBM can identify children who would require intervention later in schooling (VanDerHeyden et al, 2006). There is also an apparent overlap among cognitive competencies and constructs that early numeracy measures assess (Gersten et al, 2005). It is unclear at this time, however, if the TEN measures are able to identify children with specific learning disabilities in comparison to children who have not received adequate instruction.

Finally, in order to better understand the relationship between TEN performance and academic and behavior outcomes, further research should explore relationships with additional types of outcomes measures. In particular, future research should consider using broader academic outcome measures that assess mathematics concepts and applications, in addition to computation skills. In addition, it is still of interest to determine the link between TEN performance and behavior. Future research might use a more comprehensive measure such as a standardized behavior rating scale, for instance

the Behavior Assessment System for Children (BASC-2; Reynolds & Kamphaus, 2004) or the Child Behavior Checklist (CBCL; Achenbach, 1991), to help assess the link between early math skills and behavior problems. ODRs may be a more useful variable to examine as students reach the upper elementary or middle and high school years; however, this study and that by McIntosh et al. (2006) suggested that ODRs are limited at lower grade levels due to floor effects. That is, few students receive disciplinary referrals at this age.

Overall, future research should continue to aim to firmly establish which kindergarten measures are the best early indicators of mathematics difficulties, thereby providing schools with decision making tools for identifying students at-risk for mathematics difficulties and for monitoring student progress within an RTI framework.

Conclusions and Implications

The results of the current study indicate that TEN measures continue to demonstrate promise as indicators of early numeracy skills and may be considered early indicators of mathematics skills. Previous research demonstrated alternate-form, test-retest, and interscorer reliability as well as concurrent validity among the measures, construct validity with other measures of number sense, and predictive validity over one school year for both kindergarten and first grade TEN (Clarke & Shinn, 2004; Chard et al., 2005; Petreshock et al., 2005). This study extends previous work evaluating TEN by examining the sensitivity of these measures over multiple school years and exploring long-term predictive validity of these measures for math computation skills, teacher ratings, and report card grades. Current findings suggested that kindergarten TEN may be a useful predictor of various first grade outcomes, particularly continued early numeracy

skills, computation skills, teacher ratings, and report card grades. In this way, TEN may serve as useful benchmark or screening measures in an RTI framework. Many of the results of this study are encouraging and indicate several medium to large effect sizes among kindergarten TEN and first grade outcomes. If school officials choose to use these measures, there are several implications about which measures they might select as the best indicators of different first grade mathematics outcomes.

The MN measure demonstrates the most support as a single early indicator for kindergarten students. In almost all of the analyses, performance on the kindergarten MN measure showed significant correlations and was a significant predictor of first grade computation performance and teacher-determined outcomes measures. These findings were not surprising given that various authors (Geary, 2004; Gersten & Chard, 1999; Gersten et al. 2005) consider several cognitive correlates of the skill measured by the MN probe, such as visuo-spatial skills and storage of numerical information, to be critical determinants of later mathematics ability; and that some (Katzir & Pare-Blagoev, 2006) describe an internal number line as primarily responsible for number processing. However, the kindergarten QD measure also has potential as a predictor of first grade math performance, especially where teacher determined outcomes are of interest (i.e., teacher ratings on the ACES-M and report card grades). These findings correspond with and further contribute to the research of Chard et al. (2005) who demonstrated that kindergarten MN and QD performance were significant predictors of student performance on a standardized test of number knowledge. In accordance with their results, the MN and QD measures involve skills that are both considered key factors in number sense (Gersten et al. 2005). However, the fact that the MN and QD measures

related to different types of outcomes corresponds with findings that show that although counting knowledge and quantity discrimination skills are both key components of number sense, they are not well linked (Gersten et al.). Thus, educators may choose which measures to administer based on the outcome of interest.

Grade level is also an important consideration. Specifically, measures of lower level skills, such as the OC measure, may play a more important role earlier in school (i.e. preschool, kindergarten) while the QD and MN measures are more useful indicators of mathematics skills later on (i.e., first grade). A key aspect of any CBM probe is the ability to measure progress. The TEN measures demonstrated the ability to measure growth over time, although growth was slow on all measures and it could take several weeks to see a gain of a single digit depending on the measure used and age of the student. It is not clear whether rates of student progress would change if students were receiving more individualized intervention. In other words, while TEN measures may be useful as screening or benchmarking tools that correspond well to the first tier of the typical RTI framework, it is not clear whether they would be practical to monitor short-term progress in higher tiers in the service delivery process.

Appendix A-Pilot Study Consent Form

PARENTAL CONSENT FORM

My name is Stephanie Petreshock and I am a student in the Educational Psychology Ph.D. Program at the Graduate Center of the City University of New York (CUNY), and Principal Investigator of this project, entitled "The Reliability and Validity of Kindergarten Mathematics Curriculum-Based Measurement." This is a research study on a new way to measure mathematic ability in early-grade students. I would like your permission for your child to participate in this study.

The study requires that students complete two sessions of individual testing which will occur in a private location of your child's school. Students will be tested on 7 short measures of mathematics knowledge and skills. These measures range from 1 minute to approximately 10 minutes in length and will include different activities such as counting, pointing to numbers, and simple adding and subtraction. The testing sessions will occur twice during the school year and will last for approximately 30 minutes each. Your child will be alone with the experimenter during testing, however school staff will be present in the building. The testing will be done by trained school psychology doctoral students. In addition, with your permission, testing sessions will be audio recorded so that testing procedures can be monitored by the researcher. Students will receive stickers from their participation when testing is completed.

One potential risk of this study is missing class time. However, students will miss class for a short time and will be working on academics during the missed time. Every effort will be made to ensure that important class time is not missed. A second risk may be that students will be anxious during testing. This risk is minimal as many children find these brief activities fun. Children will be told that they can choose not to participate and can stop at any time. The researcher will be alert to any discomfort your child may experience, will stop testing, return your child to class, and if necessary inform the building school psychologist. The benefits of this study include contributing new assessment techniques to the education field for the early assessment of mathematics abilities, including early identification of mathematics difficulties. In addition, if you would like, we can provide feedback to you and/or your child's teacher regarding your child's performance. All information will be kept strictly confidential, and will be stored in a locked file cabinet, to which only I and my advisor will have access.

I may publish the results of the study, but your child's name, or any identifying characteristics will not be used in the publications. If you would like a copy of the study, please provide me with you address, and I will send you a copy in the future.

If you have questions about the research, you can contact me at (516) 987-8034 or spetreshock@gc.cuny.edu, or my advisor, Dr. Robin Codding at (212) 817-8292 or rcodding@gc.cuny.edu. If you have questions about your rights as a participant in this study, you can contact Kay Powell, IRB Administrator, The Graduate Center/City University of New York, (212) 817- 7525, kpowell@gc.cuny.edu.

Appendix A-Pilot Study Consent Form

Thank you for considering participation in this study. The attached form is your copy.

Please check one:
☐ I agree to allow my child, _____to participate in this study.
☐ I do not agree to allow my child, _____to participate in this study.

I agree to have the results of my child's performance reviewed by the building school psychologist and my child's teacher [circle one]: Yes No

I agree to have my child's assessment sessions audio-recorded [circle one]: Yes No

_____ _____ _____ _____
Parent/Guardian Signature Date Investigator's Signature Date

Appendix B-Consent Form

PARENTAL CONSENT FORM

Introduction

My name is Stephanie Petreshock and I am a student at the Graduate Center of the City University of New York (CUNY), and am conducting a research study entitled "The Long-Term Predictive Validity of Kindergarten Mathematics Curriculum-Based Measurement." This study may show that testing early math skills predicts later math skills. This means that we can help students improve mathematics sooner. Last year, your child participated in my study entitled, "The Reliability and Validity of Kindergarten Mathematics Curriculum-Based Measurement". I would like your permission for your child to participate in this new study which adds to the first study.

Procedures

- Like last year, in this study your child will take part in two 20 minute sessions of individual testing of math skills. These sessions will occur twice during the school year take place in a quiet area of the school. This year, the study will also require that we review your child's school records at the end of the school year.
- With your permission, I would like to audio record sessions to make sure that researchers say the same thing to each child. However, if you do not want your child to be audio recored, they may still participate in the study.

Risks and Benefits

- One potential risk of this study is missing class time. However, students will miss class for a short time and will be working on academics during the missed time. Every effort will be made to ensure that important class time is not missed.
- Another potential risk is any possible anxiety your child might experience in test situations. However, many children find these activities fun and children can choose not to participate and can stop at any time. The experimenter will be alert to any discomfort your child may experience and will stop testing and return your child to class. If necessary, the school psychologist will be on hand to assist with any potential discomfort.
- The benefits of this study include contributing new assessment techniques to the education field for the early assessment of mathematics abilities, including early identification of mathematics difficulties.
- In addition, if you would like, we can provide feedback to you and/or your child's teacher regarding your child's performance.

Voluntary Participation

- Your child's participation is voluntary. You may remove your child from the study at any time. If you wish to stop participation, please call me at (516) 987-8034. Whatever you decide will in no way affect your child's education program at school.
- In addition, your child can choose not to participate in the study and can stop at any time.

Appendix B-Consent Form

Confidentiality

- All information will be kept strictly confidential, and will be stored in a locked file cabinet, to which only I and my advisor will have access.
- I may publish the results of the study, but your child's name, or any identifying characteristics will not be used in the publications. If you would like a copy of the study, please provide me with you address, and I will send you a copy in the future.

Contact Information

If you have questions about the research, you can contact me at (516) 987-8034 or spetreshock@gc.cuny.edu, or my advisor, Dr. Georgiana Tryon at (212) 817-8293 or gtryon@gc.cuny.edu. If you have questions about your rights as a participant in this study, you can contact Kay Powell, IRB Administrator, The Graduate Center/City University of New York, (212) 817- 7525, kpowell@gc.cuny.edu.

Thank you for considering participation in this study. The attached form is your copy. Please check one:

☐ I agree to allow my child, _____to participate in this study.

☐ I do not agree to allow my child, _____to participate in this study.

I agree to have my child's assessment sessions audio-recorded [circle one]: Yes No

I agree to have my child's testing results shared with the school [circle one]: Yes No

_____ _____ _____ _____

Parent/Guardian Signature Date Investigator's Signature Date

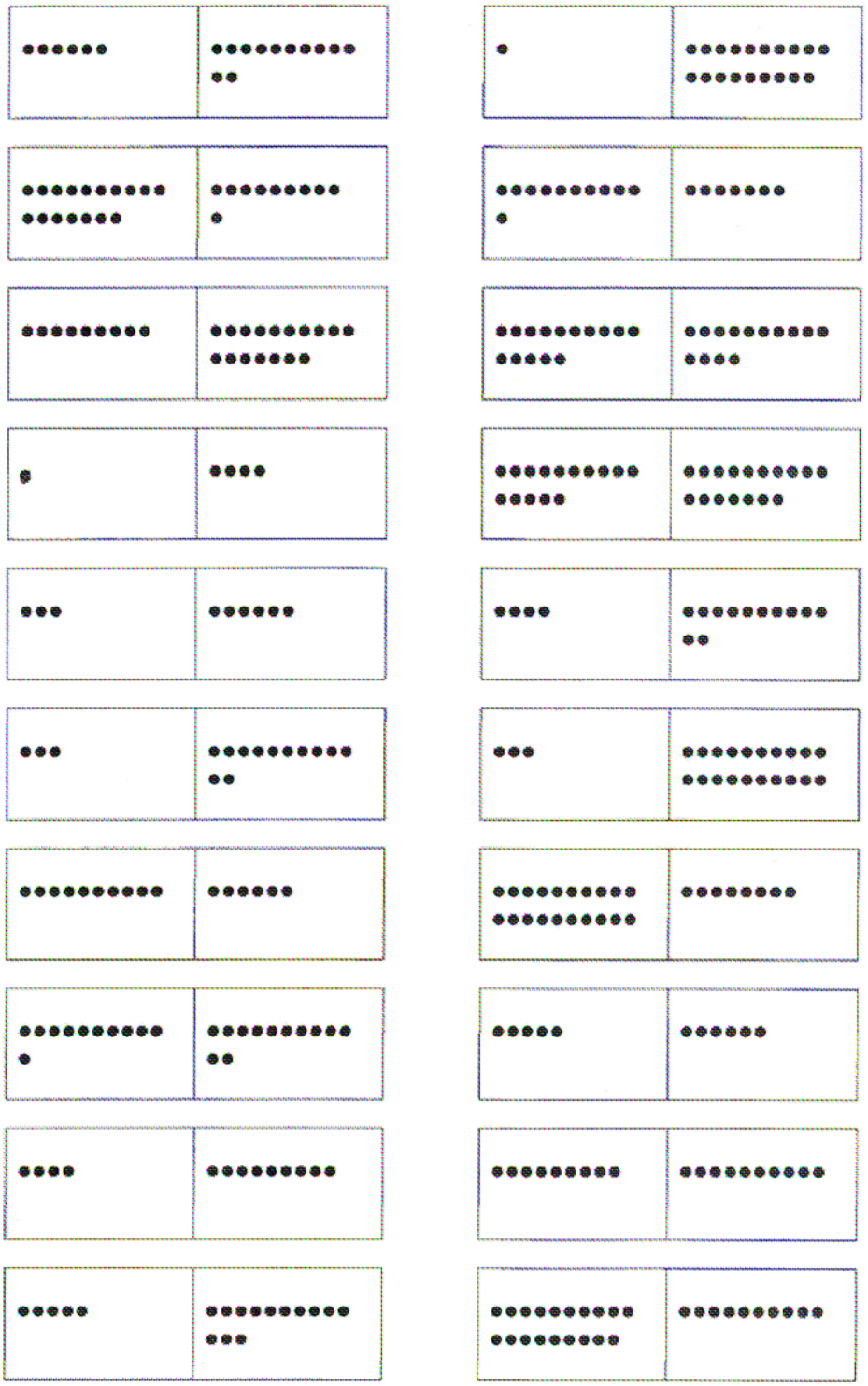
Appendix C-Attrition Group Demographics

Variable	*n*	Percent of Sample
Gender		
Male	30	60.6%
Female	13	39.4%
Ethnicity		
African-American	2	6.1%
Asian	2	6.1%
Caucasian	13	39.4%
Hispanic	16	48.5%
ESL services		
Yes	3	9.1%
No	30	90.9%
Special education services		
Yes	0	0
No	33	100.0%
Related services		
Yes	1	3.0%
No	32	97.0%

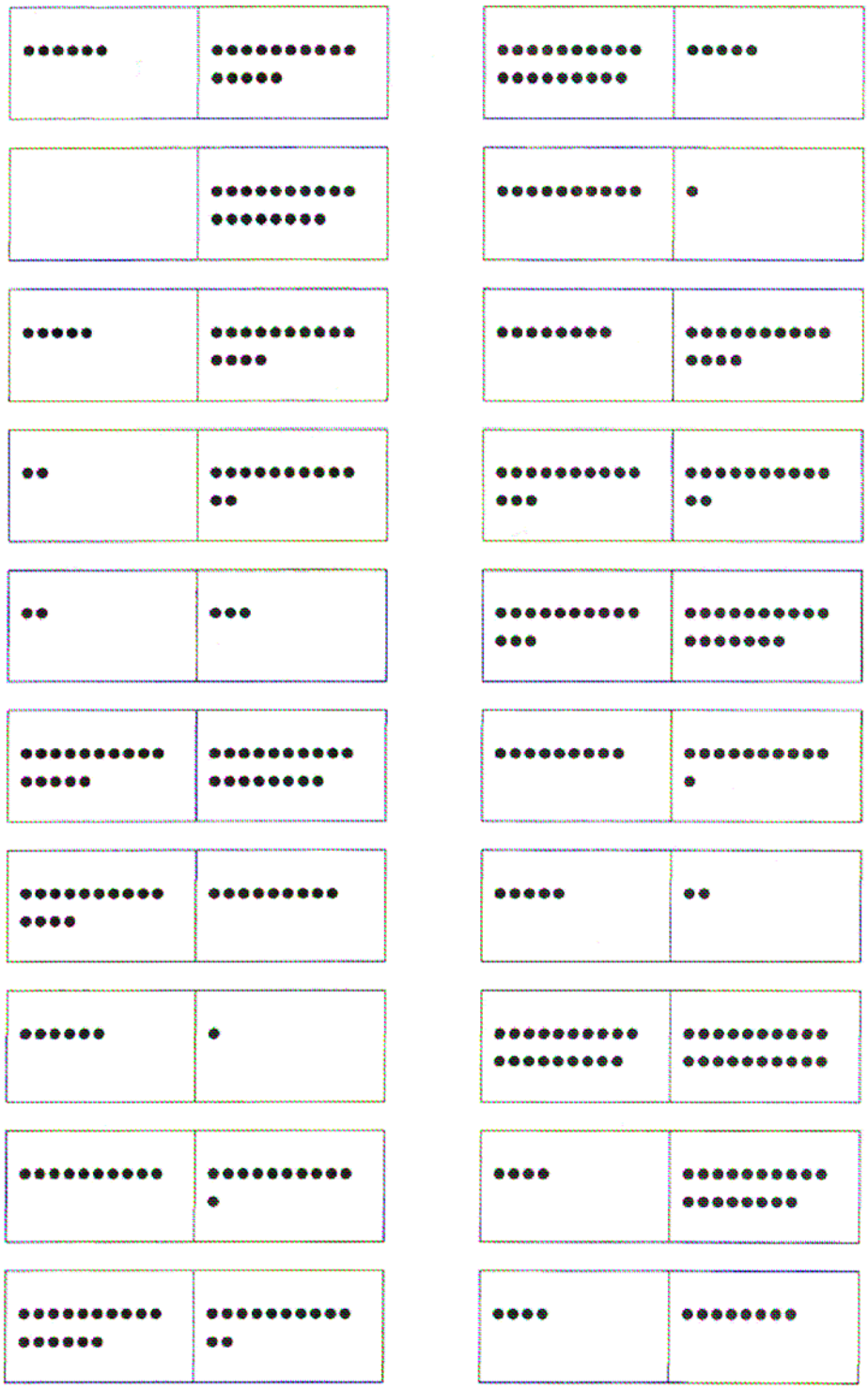
Appendix D-Participants in Each Classroom

School	Class	Number of Students
School A	1	1
	2	5
	3	4
	4	5
	5	5
School B	1	10
	2	3
	3	6
	4	2
School C	1	4
	2	5
	3	3
	4	8

Appendix E-Visual Quantity Discrimination Probe



Appendix E-Visual Quantity Discrimination Probe



Appendix F-Correspondence with AIMSweb®

Stephanie,

If it is just the probes you need, I'd suggest the AIMSweb TEN CBM Measure set:
http://www.aimsweb.com/products/cbm/en-cbm/description.php

You can order online or use the printable order form (both buttons located near the top).

Let me know what else I can provide,

Jay Anderson
Edformation, Inc.
6420 Flying Cloud Dr, STE 204
Eden Prairie, MN 55378
P: 888-944-1882 x100
F: 952-944-1884
www.AIMSweb.com

From: Gary Germann [mailto:gary.germann@edformation.com]
Sent: Friday, January 12, 2007 9:38 AM
To: Petreshock, Stephanie
Cc: Jay Anderson
Subject: Re: TEN probes

Stephanie,

I sold AIMSweb to Harcourt Assessment, Inc. For this reason it would be best if you actually subscribed to AIMSweb to secure the TEN probes. You can subscribe in several different ways but it might be best to speak directly with Jay at the office about your needs.

Gary
On Jan 9, 2007, at 7:35 PM, Petreshock, Stephanie wrote:

Hello Mr. German,
You may recall that I corresponded with you around this time last year about using AIMSweb TEN probes in a research study I was conducting in my graduate program. I am continuing my research in this area for my dissertation this year and would like to use the TEN probes in my study again. Can I obtain these measures through a subscription to AIMSweb or would you need a written request to use from me? Please let me know. Last year I obtained the measures through AIMSweb and would be happy to do so again this year, but I wanted to double check!

Much thanks,
Stephanie Petreshock
Student, Ph.D Program in Educational Psychology, CUNY Graduate Center

Appendix G- School District Letter of Cooperation

LAWRENCE
PUBLIC SCHOOLS

John T. Fitzsimons, Ph.D.
Superintendent of Schools

Elsie Friedman
Assistant Superintendent
Business

Vicki I. Karant, Ed.D.
Assistant Superintendent
Curriculum and Instruction

P.O. Box 477
Lawrence, New York 11559
http://www.lawrence.org

Telephone
Superintendent: 516/295-7030
Business: 516/295-7042
Curriculum: 516/295-7095

Facsimile
Superintendent: 516/239-7164
Business: 516/371-9250
Curriculum: 516/295-7172

December 13, 2006

Ms. Kay Powell, IRB Administrator
The Institutional Review Board at
CUNY Graduate Center

Dear Ms. Powell,

The Lawrence Public Schools has agreed for Stephanie Petreshock, a graduate student, to conduct a study entitled, "The Long Term Predictive Validity of Early Mathematics Curriculum-Based Measurement." The principals of the Number Two School, Number Five School and the Number Six School, have agreed to participate in this study.

The Lawrence Public Schools will be pleased to take part in this worthwhile study.

Very truly yours,

Vicki I. Karant, Ed.D.
Assistant Superintendent for
Curriculum & Instruction

VIK:jls

Cc: Principals

Appendix H-Institutional Review Board Permission

Office of the Vice President for Research and Sponsored Programs
Committee on the Protection of Human Subjects

The Graduate School and University Center
The City University of New York
365 Fifth Avenue
New York, NY 10016-4309
tel 212.817.7523 fax 212.817.1629

Date: Thursday, February 08, 2007

To: Petreshock, Stephanie (Educational Psychology)

Study: 06-11-1116 "The Long Term Predictive Validity of Early Mathematics Curriculum Based Measurement"

The Institutional Review Board (IRB) of The Graduate Center of the City University of New York has approved the above study involving humans as research subjects. This study was approved through expedited review based on 45CFR46.110.(a) 7.

IRB Number *(IRB #06-11-1116)* This number is an IRB number at the Graduate Center which should be used on all consent forms and correspondence.

Approval Date: *2/8/2007* **Expiration Date:** *2/7/2008*

This approval is for a one-year period. You should receive a courtesy renewal notice approximately four weeks before the expiration of this project's approval. However, it is your responsibility to insure that an application for continuing review approval has been submitted by the required time. RESEARCH MUST BE SUSPENDED IF YOUR APPROVAL HAS EXPIRED. In addition, you are required to submit a final report of findings at the completion of the project.

Consent Form: The approved and stamped consent form must be used by all subjects. You are responsible for maintaining signed consent forms for a period of at least three years after study completion.

Reporting: The principal investigator must report to the IRB any serious problem, adverse effect, or outcome that occurs with frequency or degree of severity greater than that anticipated. In addition the principal investigator must report any event or series of events that prompt the temporary or permanent suspension of a research project involving human subjects.

Modifications: All modifications of protocols involving subjects must have prior approval except those involving the prevention of immediate harm to a subject which need to be reported within 24 hours to the IRB.

If you have any questions, please do not hesitate to contact me through the IRB Office at 817-7525.

Sincerely,

Richard G. Schwartz, Ph.D.
Chair, Institutional Review Board

c: Georgiana Tryon

Please return one copy of this letter to the attention of Kay Powell at the above address:

Verification: By signing below, I acknowledge that I have received this letter and am aware of and agree to abide by all of its stipulations in order to maintain active approval status, including prompt reporting of adverse events/serious problems and annual continuing review. I am aware that it is my responsibility to be knowledgeable of all federal and state regulations including CUNY's Multiple Project Assurance (MPA) with the Department of Health and Human Services.

Stephanie Petreshock 2/20/07

Signature of Principal Investigator Date

Appendix I-Treatment Integrity Protocol

Treatment Integrity checklist

Oral Counting-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for one minute	Yes	No

Number ID-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for one minute	Yes	No

QD-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for one minute	Yes	No

MN-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for one minute	Yes	No

VQD-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for one minute	Yes	No

CBM-1

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for two minutes	Yes	No

CBM-2

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for two minutes	Yes	No

CBM-3

Says Standardized Directions	Yes	No
Says "Start"	Yes	No
Times for two minute	Yes	No

SCORE: ____/15

w/CBM:_____/24

Appendix J-Interscorer Agreement Data

Date	ID	Interscorer Agreement
3/6/2007	43	97. 1%
3/6/2007	75	97. 8%
3/6/2007	3	98.0%
3/6/2007	61	100.0%
3/6/2007	53	100.0%
3/6/2007	44	100.0%
3/6/2007	13	100.0%
3/7/2007	16	99.7%
3/7/2007	35	100.0%
3/7/2007	41	100.0%
3/8/2007	47	97.7%
3/8/2007	88	100.0%
3/8/2007	92	100.0%
3/8/2007	84	100.0%
3/8/2007	38	100.0%
3/8/2007	21	100.0%
3/15/2007	45	99.6%
3/15/2007	62	100.0%
6/11/2007	40	98.8%
6/11/2007	35	99.6%

Appendix J-Interscorer Agreement Data

Date	ID	Interscorer Agreement
6/11/2007	50	100.0%
6/11/2007	91	100.0%
6/11/2007	32	100.0%
6/11/2007	64	100.0%
6/11/2007	41	100.0%
6/11/2007	66	100.0%
6/11/2007	22	100.0%
6/12/2007	88	100.0%
6/12/2007	87	100.0%
6/12/2007	38	100.0%
6/13/2007	13	99.5%
6/13/2007	43	99.6%
6/13/2007	75	99.7%
6/13/2007	61	100.0%
6/13/2007	73	100.0%
6/13/2007	51	100.0%
6/13/2007	76	100.0%

References

Achenbach, T. M. (1991) *Integrative Guide to the 1991 CBCL/4-18, YSR, and TRF Profiles*. Burlington, VT: University of Vermont, Department of Psychology.

American Psychological Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed. text revision). Washington, DC: Author.

Batchse, G., Elliott, J., Graden, J.L., Grimes, J., Kovaleski, J.F., Prasse, D., et al. (2006). *Response to intervention: Policy considerations and implementation.* Alexandria, VA: National Association of Stated Directors of Special Education, Inc.

Berch, D.B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*, 333-339.

Bieber, G., & Choi, H. (2004). The use of curriculum-based measurement by school psychologists. *Journal of Psychological Practice, 10*, 25-36.

Brigance, A. (1985). *Brigance preschool screen*. North Billerica, MA: Curriculum Associates.

Brigance, A. (1999). *Comprehensive inventory of basic skills* (Rev. ed.). North Billerica, MA: Curriculum Associates.

Cadwell, J., & Jenkins, J. (1986). Teacher's judgments about their students: The effects of cognitive simplification strategies on the rating process. *American Educational Research Journal, 23,* 460-475.

Cawley, J.F., Parmar, R.S., Yan, W. & Miller, J.H (1998). Arithmetic computation performance of students with learning disabilities: Implications for curriculum. *Learning Disabilities Research & Practice, 13*, 68-74.

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*(2), 3-14.

Clarke, B., & Shinn, M.R. (2002). *Tests of early numeracy measures (TEN): Administration and scoring of AIMSweb early numeracy measures for use with AIMSweb*. Eden Prairie, MN: Edformation Inc.

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

Cohen, J. (1992). A Power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Daly, E.J., Wright, J.A., Kelly, S.Q., & Martens, B. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly, 12,* 268- 280.

Demaray, M.K. & Elliott, S.N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*, 8-24.

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 36*(2), 5-10.

Deno, S.L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*, 184-192.

Derr-Minneci, T.F. & Shapiro, E. (1992). Validating curriculum-based measurement in reading from a behavioral perspective. *School Psychology Quarterly, 7*, 2-16.

Dev, P.C., Doyle, B.A., & Valente, B. (2002). Labels needn't stick: "At-Risk" first graders rescued with appropriate intervention. *Journal of Education for Students Placed At Risk, 7*, 327-332.

DiPerna, J. C., & Elliott, S. N. (1999). Development and validation of the Academic Competence Evaluation Scales. *Journal of Psychoeducational Assessment*, *17*, 207-225.

Dobbs, J., Doctoroff, G.L., Fisher, P.H., & Arnold, D.H. (2006). The association between preschool children's socio-emotional functioning and their mathematical skills. *Journal of Applied Developmental Psychology, 27*, 97-108.

Dossey, J.A., McCrone, S.A., & O'Sullivan, C. (2006). Problem solving in the PISA and TIMSS 2003 Assessments (NCES 2007-029). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved December 26, 2006 from https://nces.ed.gov/pubsearch.

Dowker, A.D. (2001). Numeracy recovery: A pilot scheme for early intervention with young children with numeracy difficulties. *Support for Learning*, *16*, 6-10.

Dowker, A.D. (2003). Interventions in numeracy: Individualized approaches. In I. Thompson (Ed.), *Enhancing primary mathematics teaching* (pp.127-138). Maidenhead, UK: Open University Press.

Dowker, A. (2005). Early identification and intervention for students with mathematics difficulties. *Journal of Learning Disabilities, 38*, 324-332.

Edformation (2002). *Tests of early numeracy.* Eden Prairie, MN: Edformation Inc.

Fletcher, J.M., Coulter, W.A., Reschly, D.J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia, 54*, 304-331.

Fletcher, J.M., Shaywitz, S.E., Shankwiler, D.P., Katz, L., Liberman, I.Y., Stuebing, K.K., France, D.J., Fowler, A.E., & Shaywitz, B.A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology 86*, 6-23.

Floyd, R.G., Hojnoski, R., & Key, J. (2006). Preliminary evidence of the technical adequacy of the preschool early numeracy indicators. *School Psychology Review, 35*, 627-644.

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

Fuchs, D., & Fuchs, L.S. (2001). Responsiveness to intervention: A Blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38*, 57-61.

Fuchs, D., Mock, D., Morgan, P.L, & Young, C.L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice 18*, 157-171.

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

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

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

Fuchs, L. S., Fuchs**,** D. & Hamlett, C. L. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19,* 6-22.

Fuchs, L.S., Fuchs, D., Hamlett, C.L. & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children, 58*, 436-450.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617–641.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kupek, P., et al. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique, 19*, 23–49.

Fuchs, L.S., Fuchs, D., Hamlett, C.L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27-48.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4-15.

Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33*, 18-82.

Gersten, R., Clarke, B.S., & Jordan, N.C. (2007). *Screening for mathematics difficulties in K-3 students.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gersten, R., Jordan, N.C., & Flojo, J.R. (2005). Early identification and interventions With mathematics difficulties. *Journal of Learning Disabilities, 38*, 293-304.

Ginsburg, H.P., & Baroody, A.J. (1990). *Test of early mathematics ability* (2nd ed.). Austin, TX: Pro-ed.

Gonzales, P., Guzman, J.C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Wiliams, T. (2004). Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003. *Education Statistics Quarterly, 6,* 7-19.

Good, T. L., & Brophy, J. R. (1986). School effects. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Griffin, S. (2004). Building number sense with number worlds: A mathematics program for young children. *Early Childhood Research Quarterly, 19*, 173-180.

Griffin, S.A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk of school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp.24-49). Cambridge, MA: MIT Press.

Hale, J.B. & Fiorello, C.A. (2004). *School neuropsychology: A practitioner's handbook*. New York, NY: Guilford Press.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test* (9th ed.). San Antonio, TX: Harcourt Assessment.

Harcourt Brace Educational Measurement. (2000). *Metropolitan Achievement Test* (8th ed.). San Antonio, TX: Harcourt Assessment.

Hemingway, Z., Hemingway, P., Hutchinson, N. L., & Kuhns, N. A. (1987). Effects of student characteristics on teachers' decisions and teachers' awareness of these effects. *Journal of Special Education, 11,* 313-326.

Hinshaw, S.P. (1992). Externalizing behavior problems and academic underachievement In childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin*, *111*, 127-155.

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1400 (1997, 2004).

Jordan, N.C., Kaplan, D., Locuniak, M.N., & Ramineni, C. (2007). Predicting first grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22*, 36-46.

Kalchman, M., Moss, J., & Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. In S. Carver & D. Klahr (Eds.), *Cognition and instruction; Twenty-five years of progress* (pp.1-38). Mahwah, NJ: Erlbaum.

Karlsen, B., & Gardner, E. F. (1995). *Stanford Diagnostic Reading Test* (4th ed.). San Antonio, TX: Harcourt Assessment.

Katzir, T. & Pare-Blagoev, J. (2006). Applying cognitive neuroscience research to education: The case of literacy. *Educational Psychologist, 41*, 53-74.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement-Brief Form*. Circle Pines, MN: American Guidance Service.

Kavale, K.A., & Reese, J.H. (1992). The character of learning disabilities: An Iowa profile. *Learning Disability Quarterly, 15*, 74-94.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., & Jocelyn, L. (2004). International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective. *Education Statistics Quarterly, 6*, 20-25.

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

Mazzocco, M.M., & Thomson, R.E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*, 142-155.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside

McIntosh, K., Horner, R.H., Chard, D.J., Boland, J.B.,& Good III, R.H. (2006). The use of reading and behavior screening measures to predict nonresponse to school-wide positive behavior support: A longitudinal study. *School Psychology Review, 35*, 275-291.

Missall, K., Reschly, A., Betts, J., McConnell, S., Heistad, D., Pickart, M. Sheran, C., & Marston, D. (2007). Examination of the predictive validity of preschool early literacy skills. *School Psychology Review, 36*, 433-452.

National Assessment of Educational Progress (2005). *Mathematics Assessment*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

National Association of School Psychologists (2007). *NASP position statement on identification of students with specific learning disabilities.* Retrieved April 9, 2008, from http://www.nasponline.org/about_nasp/positionpapers/SLDPosition_2007.pdf.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematic.* Retrieved April 9, 2008, from http://standards.nctm.org/document/appendix/numb.htm.

National Joint Committee on Learning Disabilities (2005). Responsiveness to intervention and learning disabilities. *Learning Disability Quarterly, 28*, 249-260.

New York State Education Department (2006). *The New York State report card: Accountability and overview report 2005-2006*. Retrieved April 9, 2008, from https://www.nystart.gov/publicweb-rc/2006/AOR-2006-280215030000.pdf.

New York State Learning Standards for Mathematics (2005, March 15). Retrieved November 20, 2006, from http://www.emsc.nysed.gov/3-8/MathCore.pdf.

No Child Left Behind Act of 2001, Pub. L. No. 107-110. Retrieved November 12, 2006, from http://www.ed.gov/policy/elsec/leg/esea02/index.html.

Norusis, M.J. (2008). SPSS 14.0 *Advanced statistical procedures companion*. Englewood Cliffs, NJ: Prentice Hall.

Okamato, 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.

Patton, J.R., Cronin, M.E., Bassett, D.S, & Koppel, A.E. (1997). A life skills approach to mathematics instruction: Preparing students with learning disabilities for the real-life demands of adulthood. *Journal of Learning Disabilities, 36*, 178-187.

Petreshock, S., Codding, R., Johnson, M., Russo, M., Schaffer, A. (2006). *The reliability and validity of kindergarten mathematics curriculum-based measurement: An extension of earlier findings*. Manuscript submitted for publication.

Piaget, J. & Szeminska, A. (1960). *The child's conception of number*. (E.A. Lunzer, Trans.). London: Routledge and K. Paul. (Original work published 1941)

Poncy, B.C., Skinner, C.H., & Jaspers, K.E. (2007). Evaluating and comparing interventions designed to enhance math fact accuracy and fluency: Cover, copy, and compare versus taped problems. *Journal of Behavioral Education, 16*, 27-37.

President's Commission on Excellence in Special Education (2002). A new era: Revitalising special education for children and their families. Retrieved May 1, 2005, from http://www.ed.gov/inits/commissionsboards/whspecialeducation/reports/index.html.

Reynolds, C. & Kamphaus, R. (2004). *Behavior Assessment System for Children* (2nd ed.). Bloomington, MN: Pearson Assessments.

Salvia, J. & Ysseldyke, J.E. (1998). *Assessment* (7th ed.) Boston: Houghton Mifflin.

Salvia, J.A. & Ysseldyke, J.E. (2001). *Assessment in special and remedial education* (8th ed.). Boston: Houghton Mifflin.

Shapiro, E.S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York: The Guilford Press.

Shapiro, E.S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention, 30*(2), 25-32.

Shapiro, E.S., Keller, M.A., Lutz, J.G., Santoro, L.E., & Hintze, J.M. (2006). Curriculum-based measures and performance on state assessment and standardized tests. *Journal of Psychoeducational Assessment, 24*, 19-35.

Shinn, M.R. (1989). *Curriculum-based measurement: Assessing special children*. New York: The Guilford Press.

Shinn, M.R. (Ed.) (1998). *Advanced applications of curriculum-based measurement.* New York: The Guilford Press.

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* (pp. 1-32). New York: Guilford Press.

Shinn, M. & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *Remedial and Special Education, 6*, 31-38.

Skiba, R., Magnusson, D., Marston, D., & Erickson, K. (1986). *The assessment of Mathematics performance in special education: Achievement tests, proficiency tests, or formative evaluation?* Minneapolis, MN: Special Services, Minneapolis Public Schools.

Stecker, P.M. & Fuchs, L.S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*, 128-134.

Thurber, R.S., Shinn, M.R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*, 498-513.

Tindal, G., Germann, G., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Report No. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

VanDerHeyden, A.M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. *Journal of School Psychology, 44*, 533-553.

VanDerHeyden, A.M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27*, 27-41.

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., & Johnson, M.B. (1989*). Woodcock Johnson tests of achievement: Standardand supplemental batteries*. Allen, TX: DLM Resources.

Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Wright, R., Martland, J., & Stafford, A. (2000). *Early numeracy: Assessment for teaching And Intervention*. London: Chapman.