Investigating the Predictive Validity of Three Measures of Number Sense

INVESTIGATING THE PREDICTIVE VALIDITY OF THREE MEASURES OF NUMBER SENSE

A Dissertation Presented

by

BETHANY C. POLITYLO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2015

School Psychology

INVESTIGATING THE PREDICTIVE VALIDITY OF

THREE MEASURES OF NUMBER SENSE

A Dissertation Presented

by

BETHANY C. POLITYLO

Approved as to style and content by:

_____
Amanda M. Marcotte, Chair

_____
Erik W. Cheries, Member

_____
John M. Hintze, Member

_____
Craig S. Wells, Member

_____
Christine B. McCormick, Dean
College of Education

DEDICATION

For my parents, whose faith in me has never wavered, and for my Papa,

whose constant interest in my studies has always meant a great deal to me.

My ‘book,’ Papa, is finally finished.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor and committee chair, Amanda Marcotte, for her constant guidance and mentorship over the last several years. I truly cannot thank her enough for giving me the confidence to take on this project and the motivation to help me to complete it. Thanks are also due to Craig Wells for both his statistical expertise and his patience and encouragement throughout this process. I am also grateful to John Hintze for sharing his wealth of knowledge with me during the initial stages of this project, and for his guidance throughout my graduate career. Lastly, thanks to Erik Cheries for his thoughtful feedback and for opening my eyes to the world of number sense through the lens of cognitive science.

Many, many thanks go to the administrators, teachers, and students who so willingly participated in this project. Without them, this work would not have been possible. Furthermore, thank you to the several generous individuals who selflessly helped with data collection: Kelly Peneston, Karen Regis, Elizabeth Barker, Brooke Dewitt, Cheyne Levesseur, Nancy To, Bobby Storey, and Mac Furey. Special thanks are also due to Shannon Barry and Becky Allen-Oleet, not only for their help collecting data, but also for their steadfast friendship and unfailing support over the past several years.

To close, I would like to sincerely thank my friends and family – both old and new – for the love, support, and encouragement they have shown me throughout this entire process. Thanks, especially, to my parents for allowing me to pursue my own path and for always believing I would be successful. Finally, many heartfelt thanks go to my husband, Mike, for his unwavering support, admirable patience, and unconditional love. This journey has been so much more enjoyable with such a wonderful partner by my side.

ABSTRACT

INVESTIGATING THE PREDICTIVE VALIDITY OF

THREE MEASURES OF NUMBER SENSE

May 2015

BETHANY C. POLITYLO, B.S., UTICA COLLEGE

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Amanda M. Marcotte, Ph.D.

Number sense has been identified as an important foundational skill in the development of later mathematics competence. Although number sense has historically been difficult to define in the educational literature, operational definitions of the construct typically consist of a collection of early numeracy skills or "number sense components" such as quantity discrimination, rote counting, and one-to-one correspondence. Consequently, assessments of number sense tend to measure a wide variety of these skills. The purpose of this study was to investigate the predictive validity of three measures of number sense: the Test of Early Numeracy (TEN), Number Sense Brief Screener (NSB), and Early Numeracy Test (ENT). This study also sought to identify which measure or combination of measures best predicted later mathematics achievement, as measured by the Test of Early Mathematics Ability, Third Edition (TEMA-3). Number sense assessments were administered to participants at kindergarten entry and the TEMA-3 was administered at the end of kindergarten. Data were analyzed using simple linear regression analyses, multiple regression analyses, and a procedure for comparing dependent correlations. Evidence for the predictive validity of each number sense measure was demonstrated;

however, statistically, no number sense measure emerged as the best predictor of later mathematics achievement. The combination of the NSB with either the TEN or the ENT explained variation in TEMA-3 scores better than the NSB alone, but this finding may not be of clinical importance. The concurrent and predictive validities of teacher rating of student number sense were also examined. Results indicated that the TEN, NSB, and ENT all predicted TEMA-3 scores better than teacher rating of student number sense in the fall. Teacher rating of student number sense in the spring explained 42% of variation in TEMA-3 scores. Implications for practice and directions for future research are discussed.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

CHAPTER 1

STATEMENT OF THE PROBLEM

Mathematics performance in the United States has suffered over the last several decades. International statistics show that students in the U.S. perform well below that of other countries on mathematics achievement tests, and in addition, national statistics reveal that U.S. students are not currently making significant gains in mathematics (Kelly et al., 2013; National Center for Education Statistics [NCES], 2013; Provasnik et al., 2012). Although this poor performance is discouraging, it has served to draw attention toward the field of mathematics education and the importance of foundational mathematics skills. Among the foundational skills found to be critical in the attainment of mathematics achievement is the faculty of number sense, which both the National Council of Teachers of Mathematics (2000) and the National Mathematics Advisory Panel (2008) have highlighted as a vital prerequisite to later success in mathematics. A well-developed number sense allows students to understand number facts and algorithms more quickly, recognize errors, and ultimately perform mathematical computations with greater ease. Several measures have been developed to assess the construct of number sense, and this study examined the predictive validity of three of the more widely used measures of number sense: the Test of Early Numeracy (TEN; Clarke & Shinn, 2004b), the Number Sense Brief Screener (NSB; Jordan, Glutting, & Ramineni, 2008), and the Early Numeracy Test (ENT; Van Luit & Van de Rijt, 2005).

Current State of Mathematics in the United States

For the past several decades, the field of education has primarily been dominated by the research and promotion of all aspects of reading and literacy. Researchers have

worked to identify the foundational skills predictive of later reading achievement, ways to assess those necessary skills, and ways to screen for students at risk for later reading failure (National Center on Response to Intervention, 2014; National Institute of Child Health and Human Development, 2000). While this unwavering dedication to literacy is certainly of great importance, equivalent attention and resources need to be allocated to the study of mathematics. In fact, mathematics researchers would likely benefit from the extensive research on reading and literacy, as this work could serve as a useful framework for mathematics researchers to focus on and examine effective mathematics practice. This necessary focus on mathematics, however, has not occurred. As a result, the United States is slowly beginning to lose its "peerless mathematical prowess" that was once present during the twentieth century (National Mathematics Advisory Panel [NMAP], 2008, p. xi). International statistics clearly show that the U.S. performs well below that of other countries on assessments of mathematics achievement, and in addition, national statistical trends reveal that U.S. students are not currently making significant gains in mathematics (Kelly et al., 2013; NCES, 2013; Provasnik et al., 2012).

Performance of U.S. students on the 2011 Trends in Mathematics and Science Study (TIMSS), which is designed to assess fourth and eighth graders from several different countries in mathematics and science, was less than adequate (Provasnik et al., 2012). In fourth grade, eight of fifty-seven participating countries outperformed the U.S. in mathematics, and U.S. performance was not measurably different from the performance of six other countries. Eleven of fifty-six countries outperformed U.S. eighth graders in mathematics, while U.S. performance was not significantly different than the performance of twelve other countries. In addition, the average performance of

U.S. eighth graders on the TIMSS in 2011 was not significantly different than average performance in 2007 (Provasnik et al., 2012). Results from the 2012 Programme for International Student Assessment (PISA), which is an international test designed to assess the achievement of fifteen-year-olds in mathematics, science, and reading, revealed that U.S. students continue to struggle in mathematics when they reach high school. U.S. performance on the mathematics portion of the PISA was significantly lower than the average performance of international students. In fact, results indicated that students from twenty-seven of the PISA's sixty-five participating countries outperformed U.S. students in the area of mathematics (Kelly et al., 2013). Moreover, only nine percent of U.S. students performed at the top level of proficiency in mathematics on the PISA (in contrast, fifty-five percent of students from Shanghai, China performed at the top level) (Kelly et al., 2013). U.S. performance on national mathematics assessments is no less concerning. According to the National Assessment of Educational Progress, fourth grade students made no significant gains in mathematics between 2007 and 2009, and have only made small gains since 2009 (NCES, 2013). In 2013, only 42% of students were considered proficient in fourth grade mathematics, and in eighth grade, only 35% of students were found to be proficient (NCES, 2013).

Although this unsatisfactory performance is discouraging, it has served to draw much-needed attention towards the field of mathematics education. Much like what has already been done with reading, researchers are now beginning to investigate the early foundational skills necessary for children to be proficient in mathematics. These skills, as well as the importance of success in mathematics, were highlighted in 2008 as part of the Final Report of the National Mathematics Advisory Panel.

Importance of Success in Mathematics

In response to the U.S.'s increasingly poor performance in the area of mathematics, President George W. Bush called for the creation of the National Mathematics Advisory Panel (NMAP) in 2006. Members of this panel were instructed to recommend ways to "foster greater knowledge of and improved performance in mathematics among American students" while using the "best available scientific evidence" (NMAP, 2008, p. xiii). Not only did the members of the panel make recommendations regarding the foundational skills, curricula, instructional practices, and teacher education necessary for proficiency in mathematics, they also emphasized the general importance of mathematics to future success in education and in life.

Competence in mathematics is vital on both a national and individual level. Nationally, retirements and job growth in science, technology, engineering, and mathematics (STEM) fields are resulting in an increased demand for individuals with expertise in those areas (NMAP, 2008). Instead of meeting that demand domestically, however, the U.S. is relying more and more on the skills of international scientists and engineers. In fact, fewer and fewer U.S. citizens are earning degrees in STEM areas, and in turn, the U.S. is failing to produce enough individuals needed to fill the jobs available in the fields of science and technology. According to the NMAP (2008), this outsourcing of jobs and dependence on the talent of other countries has threatened the U.S.'s economic security, as well as its role as a world leader. Furthermore, the lack of U.S. independence in various STEM fields has most certainly led and will continue to lead to the slowing of this nation's technological advances (NMAP, 2008).

Individual success in mathematics is equally as important as success at the national level. Those who complete higher-level mathematics in high school (i.e., Algebra II and beyond) are more likely to go to college, graduate from college, and excel in the workplace. For example, the majority of individuals who earn over $40,000 annually have completed Algebra II or higher in high school (NMAP, 2008). Students who enroll in higher-level mathematics courses in high school are more likely to attend a four-year college, and those who complete Algebra II in high school are more than twice as likely to graduate from college (NMAP, 2008). Graduation from college, of course, leads to better job opportunities, benefits, and salaries. Clearly, individual expertise in mathematics is of great importance, as it not only leads to personal benefits but will also likely lead to success in mathematics and other STEM fields at the national level.

Defining and Measuring Number Sense

Due to the relationship between the completion of Algebra II and later success in college and in the workplace, the primary focus of the National Mathematics Advisory Panel (2008) was to identify and recommend methods by which educators and policymakers could prepare students for entry into and success in high school algebra. Members of the panel also recognized that mathematics is a hierarchical subject area, where complex skills often build on simpler, more foundational skills. In order for students to be proficient in a higher-level subject like algebra, they must first master early arithmetic skills such as the understanding of numbers, fractions, operations, and measurement (National Council of Teachers of Mathematics [NCTM], 2000; NMAP, 2008). Many of these foundational skills – counting, knowledge of whole numbers, the ability to compare quantities, and fluency in basic computations – are taught and acquired

as early as preschool and mastered throughout the early elementary grades. Perhaps most prominent among these foundational skills is the critical, yet abstract, faculty of number sense.

Both the National Council of Teachers of Mathematics and the National Mathematics Advisory Panel have highlighted number sense as a vital prerequisite to success in mathematics (NCTM, 2000; NMAP, 2008). Students who possess a strong sense of number are often better able to understand numbers and their relationships, use a variety of problem-solving strategies, recognize errors or impossible solutions, and ultimately perform mathematical computations with greater ease.

For as much as number sense is touted as a critical skill, it has historically been poorly defined in the literature (Berch, 2005; Gersten, Jordan, & Flojo, 2005; Mcintosh, Reys, & Reys, 1992). Educational definitions of number sense are often vague and overly extensive, containing abstract principles and far too many components. As Gersten et al. (2005) have noted, number sense is a complex, intricate set of skills that "no two researchers have defined in precisely the same fashion" (p. 296). In fact, Berch's (2005) brief review of the literature found approximately thirty different definitions of number sense, ranging from the ability to estimate, to the understanding of number meanings, to the skill of having a non-algorithmic "feel" for numbers. The NMAP (2008) defines number sense – among other things – as the ability to subitize, count, estimate, work with whole numbers and fractions intuitively, understand basic operations, and problem solve. A more recent, comprehensive review of the number sense literature found forty studies containing thirty-four different proposed components of number

sense, including the ability to compare quantities, identify numbers, and count objects (Politylo, White, & Marcotte, 2011).

While the vague and extensive definitions for number sense make it difficult to research, there is a growing consensus that the best way to observe and measure the construct is through the assessment of early numeracy skills (Chard et al., 2005; Clarke, Baker, Smolkowski, & Chard, 2008; Gersten & Chard, 1999; Lembke & Foegen, 2009). Much like phonemic awareness has been operationalized through the measurement of skills such as blending and segmenting the sounds of oral language, number sense has begun to be operationalized through the assessment of skills and behaviors that may represent the latent construct (Gersten & Chard, 1999; Methe et al., 2011). These early numeracy skills include counting, one-to-one correspondence, cardinality, number identification, and other similar skills acquired before, during, and just after kindergarten. In many cases, early numeracy skills are identical to the several previously identified components of number sense, such as the ability to compare quantities, estimate, and manipulate numbers. This significant overlap has led some researchers to view early numeracy skills and number sense interchangeably (Berch, 2005; Methe et al., 2011). Consequently, as will be seen in the following section, the number sense assessments that currently exist are essentially measures of early numeracy skills.

Over fifteen assessments of number sense are currently used in both research and practice, most of which operationalize number sense through the measurement of early numeracy skills. Some of these measures include the Number Knowledge Test (NKT; Okamoto & Case, 1996), the Number Sense Test (Malofeeva, Day, Saco, Young, & Ciancio, 2004), and the easyCBM measures (Alonzo, Tindal, Ulmer, & Glasgow, 2006).

As noted, many of these measures involve the assessment of early numeracy skills. For example, the NKT is a test designed to measure children's understanding of whole numbers and related concepts. Skills assessed include rote counting, one-to-one correspondence, quantity comparison, simple computation, and more complex computation (Okamoto & Case, 1996). The Number Sense Test is a similar measure of number sense created for use with preschool students, ages three through five. This assessment includes six scales designed to evaluate students' skills in the areas of counting, number identification, number-object correspondence, ordinality, comparison, and addition and subtraction (Malofeeva et al., 2004). The easyCBM measures are more broad-based measures of number sense; at the kindergarten and first grade levels, these measures assess skills in the areas of number, operations, measurement, geometry, and algebra (Alonzo et al., 2006; Clarke et al., 2011). Although initial research exists on each of these assessments, they appear to be less established and thus are not used frequently in research or practice. (For a more comprehensive review of these and other number sense measures, see Chapter 2).

The more popular measures of number sense, and thus those that appear more frequently in the literature, include the Test of Early Numeracy (TEN), the Number Sense Brief Screener (NSB), and the Early Numeracy Test (ENT). The TEN is a set of individually administered curriculum-based measures designed to assess the early numeracy skills of students in kindergarten and first grade. These measures consist of four fluency-based subtests that each take one minute to complete: Oral Counting, Number Identification, Quantity Discrimination, and Missing Number (Clarke & Shinn, 2004b). The NSB is a 33-item, individually administered assessment intended for use

with kindergarteners and first graders (Jordan et al., 2008). The measure includes items on one-to-one correspondence, number recognition and comparison, nonverbal calculation, and story problems. Finally, the ENT is a 40-item, individually administered assessment designed to measure the early mathematical competence of students in preschool through grade one (Van Luit & Van de Rijt, 2005). Originally developed by Dutch researchers, the assessment includes items on counting knowledge, concepts of comparison, seriation, classification, one-to-one correspondence, and general knowledge of numbers.

While research has been conducted on the TEN, NSB, and ENT individually, no studies have compared the predictive qualities of these three assessments directly. In other words, although preliminary research has been conducted on the reliability and validity of each measure in isolation, no study has compared the three assessments in an attempt to discern which, if any, is the best predictor of later mathematics achievement. Similarly, no studies have investigated whether or not certain combinations of the TEN, NSB, and ENT predict later mathematics achievement above and beyond that of just one measure. In addition, few studies have replicated the findings that currently exist on the psychometric properties of each measure of number sense.

Purpose of this Study

Given the lack of extensive research on the TEN, NSB, and ENT, the primary purpose of this study was to more closely examine the predictive utility of these assessments in several different ways. First, this study attempted to determine the predictive validity of each measure of number sense. Based on prior research, it was hypothesized that there would be a positive relationship between performance on each

number sense measure and later performance on a mathematics achievement test. This study also aimed to determine which of the three number sense measures, if any, was the best predictor of later mathematics achievement. It was hypothesized that one of these measures would emerge as the more effective test for predicting later mathematics achievement. Finally, this study attempted to ascertain if there was a particular combination of number sense measures that predicts later mathematics achievement better than any one number sense measure alone. For example, does performance on the TEN and the NSB predict later mathematics achievement above and beyond performance on the NSB alone? It was hypothesized that there are predictive measures of early numeracy that account for varying skills subsumed within the construct of number sense that can be modeled in a meaningful way so as to best predict later mathematics achievement. Identifying the measure or measures that best predict later mathematics achievement upon entry to kindergarten could give teachers a powerful tool with which to screen all students and provide targeted instruction so as to prevent later mathematics difficulties.

The final purpose of this study was to add to the literature base on the TEN, NSB, and ENT by replicating research that has already been conducted on these measures. Although a small body of research currently exists on each measure, few studies have replicated this work. Furthermore, while the ENT is popular abroad, there has been no research on the measure in the U.S. In addition to examining the predictive validity of each measure, this study also provided data to assess the inter-rater reliability of each assessment, as well as the relationship between performance on each measure and teacher rating of student number sense.

Research Questions and Hypotheses

Predictive Validity of Number Sense Measures and Teacher Rating

1. Is there a relationship between kindergarteners' fall performance on each measure of number sense and spring performance on a mathematics achievement test?
2. Which measure of number sense, administered in the fall of kindergarten, best predicts mathematics achievement in the spring of kindergarten?
3. Is there a relationship between a teacher's rating of a kindergartener's number sense in the fall and that same kindergartener's performance on a mathematics achievement test in the spring?

This first set of research questions aimed to assess the predictive validity of each measure of number sense. In addition, the predictive validity of teacher rating of number sense was examined. It was hypothesized that there would be a strong positive relationship between fall performance on each measure of number sense and spring performance on a mathematics achievement test (i.e., the Test of Early Mathematics Ability, Third Edition [TEMA-3]). Second, because the ENT is the most comprehensive measure of number sense and assesses the broadest range of early numeracy skills, it was hypothesized that the ENT would be the best predictor of later performance on the TEMA-3. Finally, it was predicted that there would be a positive relationship between a teacher's rating of a kindergartener's number sense in the fall and that same kindergartener's spring performance on the TEMA-3.

Predictive Validity of Combinations of Number Sense Measures and Teacher Rating

4. Is there a combination of number sense measures that predicts mathematics achievement above and beyond that of just one measure? For example, does the TEN

and the NSB predict mathematics achievement above and beyond that of the NSB alone?

5. Does performance on a number sense measure, combined with teacher rating of number sense, predict mathematics achievement above and beyond that of performance on a number sense measure alone?

The next group of research questions examined the predictive validity of different combinations of number sense measures. In addition, the predictive validity of a number sense measure plus teacher rating was examined. It was hypothesized that all combinations of number sense measures created would predict mathematics achievement above and beyond that of just one measure alone. In other words, it was predicted that all three number sense measures would predict mathematics achievement above and beyond that of any one measure alone; similarly, any combination of two number sense measures would predict mathematics achievement above and beyond that of any one measure. These hypotheses were developed due to the fact that all three measures of number sense, while assessing the same construct, all contain at least a couple of unique items which measure different early numeracy skills (e.g., the NSB is the only measure that assesses counting principles, and the TEN is the only measure that requires the student to identify the missing number in a sequence of three digits). Finally, it was hypothesized that performance on a number sense measure combined with fall teacher rating of number sense would predict mathematics achievement above and beyond that of performance on a number sense measure alone.

Concurrent Validity of Teacher Rating

6. Is there a relationship between a teacher's rating of a kindergartener's number sense in the spring and that same kindergartener's mathematics achievement in the spring?

The final research question assessed the concurrent validity of a teacher's rating of student number sense. It was hypothesized that there would be a positive relationship between a teacher's rating of a kindergartner's number sense in the spring and that same kindergartener's spring performance on the TEMA-3.

CHAPTER 2

LITERATURE REVIEW

Historical Foundations of Number Sense

> Man, even in the lower stages of development, possesses a faculty which, for want of a better name, I shall call *Number Sense*. This faculty permits him to recognize that something has changed in a small collection when, without his direct knowledge, an object has been removed from or added to the collection. (Dantzig, 1946, p. 1)

Although the term was only first used by mathematician Tobias Dantzig in the early twentieth century, there is evidence that "number sense" has existed for hundreds of thousands of years – long before numbers or numerals were ever documented or communicated through oral, symbolic, or written language. As Dantzig (1946) notes, number sense is historically described as the innate ability to perceive small, concrete quantities without counting. This faculty also allows for the ability to identify the difference in size between two small groups and to recognize when an element has been removed from or added to a group. Given the intuitive nature of number sense, it is impossible to pinpoint exactly when the faculty developed. Research conducted with a wide variety of animal species, however, suggests that number sense existed well before humans walked the planet (Dehaene, 1997; Ifrah, 1985; Rilling, 1993).

History brings with it many anecdotes describing the apparent mathematical prowess of various animals. Perhaps the most familiar story is that of Clever Hans, a horse who could seemingly solve both simple and complex mathematical problems with the tapping of his hoof (Fernald, 1984). Hans' feats garnered much attention and skepticism during the early 1900s; his demonstrations in Germany were frequent and spectators often eagerly gathered to watch Hans flawlessly solve computation problems,

tell time, and add fractions. Of course, a series of rigorous experiments by psychology graduate student Oskar Pfungst eventually demystified Hans' abilities (Fernald, 1984). Hans was not, in fact, able to add, subtract, or solve any kind of mathematics problems on his own. His mathematical intelligence was simply a reflection of his owner's knowledge, as when Hans arrived at the correct answer to a particular problem, his owner would unconsciously tilt his head back ever so slightly. Hans was sensitive to this small movement and he learned to stop tapping his hoof when he saw this signal, thereby making it appear that he had solved the problem on his own (Fernald, 1984).

While the Clever Hans phenomenon left the scientific community suspicious of animal intelligence for many years, it also fueled a strong interest in researching the true mathematical abilities of animals (Dehaene, 1997). Countless studies have been conducted investigating the numerical competence of animals (Dehaene, 1997; Rilling, 1993), and several highlight the notion that animals do indeed possess an ability to recognize, manipulate, and distinguish between small quantities of number. Some of the first experimental studies that revealed animals' understanding of number were conducted by Otto Koehler in the mid 1900s. Koehler worked with birds, and his experiments demonstrated a bird's ability to recognize and discriminate between quantities (Koehler, 1951). In one of his seminal studies, his subjects – a raven named Jacob and a grey parrot named Geier – were presented with five boxes. Each box had a different number of dots printed on it, ranging from two to six. On the ground, there was also a pattern of dots ranging in quantity from two to six. Both birds learned to only open the box that showed the same number of dots that was also on the ground (Koehler, 1951). In other experiments, Koehler trained birds to eat exactly five mealworms from a

row of jars containing zero, one, two, or three mealworms. During the training phase, the birds would be shooed away after eating five of the worms. Eventually, the birds learned to eat exactly five worms and then walk away on their own (Koehler, 1951). Koehler ultimately concluded that while birds do not have the ability to count or name numbers, they do appear to demonstrate "unnamed thinking;" in other words, birds are able to construct internal representations of small quantities of number without the use of language (Koehler, 1951).

Abilities similar to those found in Koehler's birds have also been demonstrated in rats. Beginning in the 1950s, Francis Mechner, an American psychologist, began investigating the numerical competence of rats. Using Skinner boxes, Mechner (1958) trained rats to press two levers in order to receive food. Essentially, in order to receive food pellets, the rats had to press the first lever a specified number of times (e.g., either 4, 8, 12, or 16 times) and then press the second lever once. Mechner found that the rats became quite adept at pressing the first lever the approximate number of times needed. Rats who needed to press the first lever four times, generally pressed it four or five times; those who needed to press the first lever eight times also generally did so. An interesting point is that as the number of required presses increased (e.g., 12 or 16), the rats became less precise with their presses. Rats who were required to press the lever 16 times, for example, sometimes pressed the lever anywhere from 12 to 24 times (Mechner, 1958). Much like Koehler's studies showed, Mechner's findings support the notion that animals are able to approximate at least small quantities of number. Larger quantities, however, appear more difficult for animals to discriminate between and internally represent.

More recent research has shown that pigeons and rats are not the only species with a seemingly innate sense of number. Woodruff and Premack's (1981) work with chimpanzees demonstrated that chimps are not only able to recognize small, whole number quantities, but they are also able to discriminate between fractional quantities such as one quarter, one half, and three quarters. Woodruff and Premack's experiment was set up as such: a chimp was shown a sample (e.g., food, circular wooden disks, glass jars filled with blue-colored water) and would then have to select one stimulus from a set of two that correctly matched the sample. For example, the chimp might have been presented with a sample consisting of two glasses filled completely with water. Then, two stimuli would be presented: a tray with three apples and a tray with two apples. To answer the trial correctly, the chimp would have to choose the tray with two apples. Researchers found that the chimps were remarkably good at selecting the correct stimulus, even when the size, shape, and type of stimuli or sample changed. One particular chimp, Sarah, was even adept at matching fractions. When presented with a sample of a half-filled glass, for example, Sarah could consistently select the correct stimulus that showed half an apple or half a wooden disk. The same was true for the fractions one quarter and three quarters. Woodruff and Premack (1981) concluded that while chimps can recognize and discriminate between small whole numbers, they also appear to have a basic understanding for part-whole relationships and analogical reasoning.

Basic numerical competence has been demonstrated in lions, dolphins, and parrots, as well. By using playback of either one or three lions roaring, McComb, Packer, and Pusey (1994) discovered that prides of lions would often approach the playback of a

single lion roaring in order to defend their territory, but they tended not to approach the playback of three lions roaring unless the size of their pride was much larger. It seems "numerical assessment skills" such as the ability to discern the relative size of a group are helpful in contributing to the overall survival and fitness of lions and other species (McComb et al., 1994). An understanding of the relative size of different groups has also been shown in dolphins. When presented with two sets of objects, dolphins have been successfully trained to select the set with the fewest number of objects, even when the sets change in number and the objects are novel (Kilian, Yaman, Von Fersen, & Güntürkün, 2003). In addition, Kilian et al. (2003) found that dolphins could discriminate between slightly larger numbers such as five and six.

Perhaps the most advanced mathematical abilities demonstrated by an animal have come from Alex, an African Grey parrot who had been the subject of several studies in the fields of cognition and communication (Pepperberg, 2006). In Pepperberg's (2006) study, Alex, who was previously trained to label quantities up to six using the English language, showed that he could complete basic addition problems. When asked "How many nuts total?" after being shown two nuts under one cup and one nut under another cup, Alex could consistently respond with three, even when the addends changed. Overall results from the study indicated that Alex could correctly solve a great variety of addition problems with sums up to six. Pepperberg (2006) concluded that with training, animals can have expanded numerical capacities that develop much like those found in young children.

Given society's impressive advances in mathematics over recent centuries, it is hard to believe that humans' concept of number was once as primitive as that of animals.

History suggests, however, that that was indeed the case: humans' understanding of number was first limited to a type of number sense that was no more advanced than the numerical competence demonstrated by the animals previously described. Anthropological studies have illustrated that several groups of isolated, indigenous peoples are not able to represent numbers greater than three or four (Dantzig, 1946; Gordon, 2004; Ifrah, 1985; Pica, Lemer, Izard, & Dehaene, 2004). People from tribes in remote areas of Australia, Africa, Brazil, and the Torres Strait, for example, all have number words for one, two, and sometimes three or four, but no words for five, six, ten, and so on. For quantities beyond the number three or four, these people use words meaning "many" or "a multitude" (Gordon, 2004; Ifrah, 1985; Pica et al., 2004). This speaks to the notion that much like animals, humans possess a rudimentary number sense that only allows for the ability to perceive and distinguish between small quantities of number. Without the ability to count, numbers larger than three or four become, as Dehaene (1997) describes, "fuzzy" approximations.

The difficulty in perceiving numbers greater than three or four is reflected throughout history. In nearly all societies, from the Mayans to the Romans to the Chinese, the first three or four numerals have always been represented by several instances of the same symbol (Dehaene, 1997; Ifrah, 1985). In Roman societies, the numbers one, two, and three, were represented by one, two, or three bars (i.e., I, II, III). In Mayan societies, these numbers were represented by dots (i.e., •, ••, •••) (Dehaene, 1997). After the numbers three or four, however, all societies begin to represent numbers by using more abstract symbols. For the Romans, the number five was denoted by "V," and for the Mayans, five was represented by a long horizontal line. Why not simply use

five bars or dots to denote the number five? The answer once again goes back to the notion that without the ability to count, humans were unable to automatically perceive and recognize five dots as representative of the quantity "five." Any quantity greater than three or four, then, was simply understood as "many" and could only be expressed in imprecise approximations (Dehaene, 1997; Ifrah, 1985).

As Dantzig (1946) notes, this inability to precisely perceive larger numbers did not render "primitive" peoples incapable of working with and understanding larger quantities of number. He provides an excellent illustration of how concepts of equality and quantity discrimination can be understood without counting:

> We enter a hall. Before us are two collections: the seats of the auditorium, and the audience. Without counting, we can ascertain whether the two collections are equal and, if not equal, which is greater. For if every seat is taken and no man is standing, we know without counting that the two collections are equal. If every seat is taken and some in the audience are standing, we know without counting that there are more people than seats (Dantzig, 1946, pp. 6-7).

This concept of one-to-one correspondence – where one item in a set is matched to one item in another set – is one of the first documented number methods used by people who lived tens of thousands of years ago (Dantzig, 1946; Ifrah, 1985). In order to calculate the total number of sheep in a herd without counting, for example, a shepherd might make a notch in a bone for every sheep that he owns. If he wants to check to see if all of the sheep are present in the future, he simply has to match each sheep with each notch. If he finds that there are as many sheep as there are notches, he knows his whole herd is present (Ifrah, 1985). Archaeologists have found evidence of this method dating back to the Upper Paleolithic period, which occurred at least 30,000 years ago. Of course, historians agree that undocumented number systems involving the use of the body likely existed before matching or tallying methods, although it is not possible to pinpoint

exactly when those systems began. Number systems that utilize the body, which are still used by indigenous peoples such as the islanders of the Torres Strait, involved touching various parts of the body in a specific, fixed sequence in order to communicate certain quantities. To communicate what modern-day society would consider "seven," for example, one might touch all five fingers of the right hand, followed by touching the right wrist and right elbow (Ifrah, 1985). In addition to those utilizing the body and notching, other forms of ancient, concrete number systems involved the use of pebbles, clay tokens, and strings. None of these methods, however, required the use of number words, written numerals, or any abstract understanding of number whatsoever.

Eventually, with the advent of number words and then written numerals, the shift from concrete to abstract and much more complex number systems began (Dantzig, 1946; Ifrah, 1985). Although a comprehensive analysis of this shift from concrete to abstract is beyond the scope of this review, it is important to note that no matter how advanced society's number systems have become, they all began with a basic concept of number, much like that of animals, and much like the "number sense" described by Dantzig (1946). Without this primitive, seemingly innate sense of number, humans would not have been able to develop the ability to estimate, count, or calculate, nor would they have been able to accomplish the impressive mathematical advancements that have been made to date.

Modern-Day Number Sense

Does this ancient faculty of number sense exist in humans today, just as it did tens of thousands of years ago? Are humans today able to perceive small quantities of number without counting, or recognize differences between two sets of items? As

Dantzig (1946) points out, counting has become "such an integral part of our mental equipment that psychological tests on our number perception are fraught with great difficulties" (p. 4). Experiments conducted within the last few decades, however, have examined humans' basic concept of number without the confounding issue of counting. These experiments have primarily been carried out with participants who do not yet know how to count or calculate – infants.

Numerical Abilities of Infants

In 1980, psychologists Starkey and Cooper conducted one of the first truly rigorous experiments that supported the notion that humans are born with a primitive sense of number. In their study, Starkey and Cooper (1980) assessed whether infants ages 16 to 30 weeks could detect the difference between small quantities of number: two and three, and then four and six. First, during the habituation phase, two dots were shown to an infant on a screen several times. These dots changed in position on different slides. Next, the infant was shown a slide with three dots. Experimenters found that the infants fixed their gaze on the slide with three dots significantly longer than the two-dot slides they were habituated to. Consequently, researchers concluded that the infants detected a change in the quantity of the dots. To be sure infants were not gazing at the three-dot slides for a longer length of time simply because there were more items on the screen to look at, Starkey and Cooper (1980) also tested infants in the reverse direction. In other words, infants were first habituated to three dots and then shown a slide of two dots. Researchers found that the infants looked significantly longer at the two-dot slide, suggesting that they were once again able to detect a difference in the number of dots shown (Starkey & Cooper, 1980). Interestingly, the infants did not appear to detect a

change in the slides that shifted from four dots to six dots, which supports the idea that the primitive number sense does not extend beyond very small quantities such as three or four. Of note is that Starkey and Cooper (1980) did not refer to the infants' numerical abilities as number sense. Instead, they concluded that part of an infant's numerical competence is the ability to subitize, or instantly "distinguish among arrays containing fewer than four items" (Starkey & Cooper, 1980, p. 1033).

In a follow-up study, Starkey, Spelke, and Gelman (1983) investigated whether infants' early numerical abilities were purely visual, or if infants could detect relationships between sets of visual and auditory items. In this experiment, infants six to eight months old were shown two screens, one containing a set of two items and the other containing a set of three. While looking at the screens, the infants then heard either two or three drumbeats, which originated from a speaker in the middle of the two screens. What Starkey et al. (1983) found was that the infants gazed significantly longer at the two-item screen when they heard two drumbeats, and they also gazed longer at the three-item screen when hearing three drumbeats. Researchers concluded that an infants' numerical competence is not solely a visual modality (Starkey et al., 1983). More recent studies using this same paradigm have demonstrated that infants who are only a few hours old will similarly gaze at a visual array that matches an auditory sequence for significantly longer than an array that does not match the auditory stimulus (Izard, Sann, Spelke, & Streri, 2009).

Xu and Spelke (2000) later challenged the idea that infants can only understand quantities up to three or four. In their study, they showed that six-month-old infants could discriminate between larger numerosities. Employing methods previously

described, infants were habituated to slides that all showed eight dots. Then, the infants were shown a sequence of slides containing either eight dots or sixteen dots. The infants did not pay much attention to the slides with eight dots, but their gaze was fixed significantly longer on those slides showing sixteen dots (Xu & Spelke, 2000). Likewise, infants who were habituated on sixteen dots were now more interested in the slides showing eight dots. In an additional experiment, Xu and Spelke (2000) changed the ratio of dots from 1:2 (i.e., 8 dots and 16 dots) to 2:3 (i.e., 8 dots and 12 dots). Infants were unable to discriminate between sets of 8 dots and sets of 12 dots, suggesting that they can only detect differences between larger numerosities when the ratio of difference between the sets is quite large. A related study by Lipton and Spelke (2003) provided similar evidence: six-month-old infants could only discriminate between large quantities when they differed by a ratio of at least 1:2. Interestingly, though, Lipton and Spelke (2003) found that nine-month-old infants were able to discriminate between sets that differed in quantity by a ratio of 2:3, but not when they differed by a ratio of 4:5. Researchers concluded that an infant's sense of number appears to develop in precision during the first months of life, long before verbal development or formal teaching of any kind (Lipton & Spelke, 2003).

Is the numerical competence of infants limited to noticing differences between two sets of quantities? Wynn's (1992) seminal study on addition and subtraction in infants illustrated that a young person's early numerical abilities clearly extend beyond that of quantity discrimination. Participants in Wynn's study were four- to five-month-old infants placed in front of a display area to watch addition and subtraction problems concretely acted out. First, the experimenter would place a Mickey Mouse toy in the

display case. Then, a small screen would rise up to cover the toy. From the side of the screen, where the infant could still see, a second identical toy was placed behind the screen. Wynn configured the experiment to have one of two outcomes when the screen dropped: (1) the expected/possible outcome of two toys behind the screen, or (2) the unexpected/impossible outcome of one or three toys behind the screen (for the impossible outcomes, the experimenter would remove or add a toy through a trap door, out of view from the infant) (Wynn, 1992). Results indicated that the infants spent significantly more time studying the unexpected/impossible outcomes than the expected/possible outcomes. The same was true when the infants were presented with possible and impossible outcomes of a basic subtraction problem. Essentially, the infants were surprised when one plus one did not equal two, suggesting that infants have some type of mechanism that allows them to understand the processes behind simple calculations. In fact, Wynn (1992) went so far as to conclude that "humans innately possess the capacity to perform simple arithmetical calculations, which may provide the foundations for the development of further arithmetical knowledge" (p. 750).

Given the host of numerical skills present at birth, Starr, Libertus, and Brannon (2013) wanted to investigate whether the acuity of an infant's number sense – or Approximate Number System (ANS) – predicts later mathematics development. To test the acuity of the ANS, Starr et al. (2013) placed infants in front of two screens. One screen showed an array of dots that stayed constant in number, but the size and placement of the dots varied. The other screen showed an array of dots that changed in number. Infants who looked longer at the screen with the dots that changed in number were said to have greater ANS acuity, as these infants were detecting the change in the number of dots

(Starr et al., 2013). Approximately three years later, Starr et al. (2013) tested these same infants using a variety of standardized early mathematics and general intelligence assessments. Researchers found that children who performed better on the standardized assessments had better ANS acuity as infants. This was the case even after controlling for general intelligence. While Starr et al.'s (2013) study provided evidence for the relationship between primitive number sense and later mathematics development, the researchers were quick to caution that this relationship is not clear-cut or one-directional. Although ANS acuity appears to be a predictor of later mathematics development, it explains only a small proportion of the variance, suggesting that many other factors affect the mathematics achievement of young children. Additionally, Starr et al. (2013) acknowledged that the relationship between ANS acuity and mathematics development is likely bidirectional; ANS acuity may contribute to mathematics development just as much as mathematics development affects ANS acuity. Although further research is certainly needed, Starr et al.'s (2013) study did support the notion that (1) an innate number sense or Approximate Number System exists in the first years of life and (2) these innate number skills are in some way related to later mathematics development.

Number Sense and the Approximate Number System in Early Childhood

Evidence indicates that infants are born with some level of numerical competence, often in the form of the ability to distinguish between sets consisting of different quantities. As Lipton and Spelke (2003) demonstrated, this number sense or Approximate Number System (ANS) appears to become more precise over time. Nine-month-old infants, for example, can discriminate between sets that differ in quantity by a ratio of 2:3, but six-month-old infants cannot. Does number sense or the ANS continue

to develop in precision over the course of early childhood, or does this faculty cease to further develop after a certain age? Working with three- to six-year-olds, as well as with adults, Halberda and Feigenson (2008) sought to answer these questions. In their study, participants sat in front of a screen that showed, for example, the following sequence: (1) X slices of pizza followed by the phrase "Here are Big Bird's pieces of pizza;" (2) Y slices of pizza followed by the phrase "Here are Grover's pieces of pizza;" and (3) X and Y slices of pizza (on separate sides of the screen) followed by the phrase "Who has more pieces of pizza?" Each part of the sequence was shown on the screen for exactly two seconds. Participants chose which character had more pieces of pizza by pressing a yellow key for Big Bird and a blue key for Grover. Participants were given feedback on every trial. As they hypothesized, Halberda and Feigenson (2008) found that number sense in children appears to increase in precision over time, before and after formal mathematics instruction begins. Three-year-olds were able to accurately discriminate between sets differing by a ratio of 3:4, four-year-olds 4:5, and five- and six-year-olds 5:6. Adult participants were able to consistently discriminate between sets differing by a ratio of 10:11 (Halberda & Feigenson, 2008). Interestingly, a study by Cantlon and Brannon (2006) uncovered similar findings with rhesus monkeys; much like humans, rhesus monkeys were readily able to discriminate between smaller ratios but had difficulty accurately discriminating between larger ratios. The researchers concluded that the acuity of the ANS does indeed increase over time, both before and during formal mathematics instruction in school. In fact, although additional research is needed in this area, they estimated that the ANS does not fully develop until the preteen years (Halberda & Feigenson, 2008).

A similar, follow-up study by Libertus, Feigenson, and Halberda (2011) found a significant relationship between preschoolers' acuity of the ANS and their performance on an assessment of early mathematics ability. These findings still held true even when controlling for age and verbal skills (i.e., vocabulary size). While research strongly suggests there is a link between the ANS and mathematics ability, it remains unclear whether the precision of the ANS predicts mathematics performance; like Starr et al. (2013) cautioned, Libertus et al. (2011) acknowledged that the relationship is most likely bidirectional, and/or that both mathematics performance and ANS acuity are mediated by additional, unknown factors.

Piazza et al. (2010) provided further evidence of the relationship between number acuity and mathematics performance. Using methods much like those of Halberda and Feigenson (2008) in which participants had to choose which array of dots was larger without counting, Piazza et al. (2010) revealed two main findings. First, as prior research has demonstrated, the ANS becomes more precise over time. Adults were able to discriminate between arrays differing by smaller ratios while kindergarteners were only able to discriminate between arrays differing by large ratios. Second, Piazza et al. (2010) found that students who had previously been identified as dyscalculic had significantly impaired number acuity when compared to normally developing peers. In essence, students who had previously identified learning disabilities in mathematics were less adept at discriminating between two sets of arrays; these results remained true even when controlling for age, IQ, and reaction times. These findings lend additional support to the notion that the ANS, which is arguably present at birth, matures over time and is strongly related to a student's level of success with traditional mathematical concepts. Those with

a less precise number sense may be at risk for developing a later learning disability in mathematics (Piazza et al., 2010).

Taken together, research over the last several decades suggests humans are indeed born with a primitive sense of number, much like the one Dantzig (1946) described many years ago. This number sense, or ANS, allows for the detection of differences between two sets of quantities, even just hours after birth. The ANS then matures over the course of infancy, childhood, and adolescence so that more subtle differences between quantities can be detected. As several studies have demonstrated, there also appears to be a strong link between number sense acuity and later success in more traditional mathematics tasks (Halberda & Feigenson, 2008; Libertus et al., 2011; Piazza et al., 2010; Starr et al., 2013). Although it remains uncertain whether there is a direct, causal relationship between ANS acuity and mathematics achievement, the relationship between the two has potential implications for educators. If educators can assess a child's number sense early in their schooling and subsequently intervene to improve the precision of a child's ANS, then this increased precision may be one of the factors that contributes to success and achievement in mathematics later in life.

Number Sense within the Field of Education

Thus far, this review has largely focused on the definition, development, and acuity of number sense through the lens of developmental and cognitive psychology. As evidenced by the studies previously reviewed, developmental and cognitive psychologists define number sense much like it was originally understood. To these researchers, number sense is characterized by an innate understanding of approximate numerical magnitudes and involves the ability to automatically discriminate between sets comprised

of different quantities. This number sense, or Approximate Number System (ANS), develops over time and is related to an individual's later success with traditional mathematical concepts (Halberda & Feigenson, 2008; Libertus et al., 2011; Piazza et al., 2010; Starr et al., 2013). Given the link between the ANS and future performance in mathematics, combined with U.S. students' overall poor performance in mathematics as of late, interest in this intriguing faculty of number sense has recently spread from the realms of developmental and cognitive psychology to the field of education. Educators are now beginning to recognize number sense as a critical prerequisite to later success in mathematics (NCTM, 2000; NMAP, 2008). Consequently, attempts to define, measure, and teach the construct have become commonplace among researchers and practitioners in the field of education. As will be seen, however, educators have typically taken a much different approach than cognitive psychologists in defining and measuring the faculty of number sense.

Defining and Operationalizing the Construct of Number Sense

While researchers in the fields of developmental and cognitive psychology conduct their research using the original, parsimonious definition of number sense, the definitions used in the field of education are much more complex, and at times, quite nebulous. For as much as the field of education touts number sense as a critical skill, it has historically been poorly defined in the literature (Berch, 2005; Gersten et al., 2005; Mcintosh et al., 1992). Educational definitions of number sense are often vague and overly extensive, containing abstract principles and far too many components. As Gersten et al. (2005) have noted, number sense (within the field of education) is a complex, intricate set of skills that "no two researchers have defined in precisely the

same fashion" (p. 296). In fact, Berch's (2005) review of the literature revealed approximately thirty different definitions and components of number sense, ranging from the ability to estimate, to the understanding of number meanings, to having a non-algorithmic "feel" for numbers. It remains unclear as to why definitions of number sense differ so greatly among cognitive and educational researchers, although Berch (2005) hypothesizes that for educators, the expansive definitions of number sense are simply incredibly engrained in the materials, curricula, and assessments already widely used in the field. Complex definitions of number sense, for example, appear in the National Council of Teachers of Mathematics' *Principles and Standards for Mathematics*, in mathematics textbooks, and on various national and international assessments (Berch, 2005).

Although definitions of number sense vary considerably, there do appear to be two main definition types: conceptual and operational. Educational researchers who have attempted to define number sense conceptually tend to look at number sense as a broad, intuitive construct. As a result, these researchers have collectively acknowledged that when number sense is viewed conceptually, it is an intricate, abstract skill that is quite difficult to observe, measure, and even understand. Various models for breaking down the complexity of the construct have been proposed (Greeno, 1991; Mcintosh et al., 1992; Okamoto & Case, 1996), yet even with these conceptual frameworks, the vague and elusive nature of the faculty remains. Several examples of conceptual definitions of number sense exist in the literature on mathematics education. Howden (1989) wrote the following:

> Number sense can be described as good intuition about numbers and their relationships. It develops gradually as a result of exploring numbers, visualizing

> them in a variety of contexts, and relating them in ways that are not limited by traditional algorithms. (p. 11)

Similarly, Sowder (1989) described number sense as a "well-organized conceptual network that enables a person to relate number and operation properties… number sense is not a body of knowledge; rather it is a way of thinking" (p. 4). In their most recent edition of *Principles and Standards for School Mathematics*, the National Council for Teachers of Mathematics (2000) paraphrased Sowder (1992) and took a more conceptual approach in defining number sense, as well:

> Central to this Standard [of Number and Operations] is the development of number sense – the ability to decompose numbers naturally, use particular numbers like 100 or ½ as referents, use the relationships among arithmetic operations to solve problems, understand the base-ten number system, estimate, make sense of numbers, and recognize the relative and absolute magnitude of numbers. (p. 32)

Despite the variability found among conceptual definitions of number sense, they all tend to define the construct more broadly while alluding to the intuitive, complex nature of the faculty. While conceptual definitions are valuable in helping educators gain a general understanding of number sense, they are not entirely useful in helping educators and researchers to observe, assess, and teach number sense. How would one, for example, objectively measure a student's intuition about numbers and their relationships? For educational researchers and practitioners, operational definitions of number sense are much more useful in the observation and assessment of this complex, latent construct.

Similar to conceptual definitions, operational definitions of number sense differ greatly. They do, however, share a common theme: operational definitions of number sense typically consist of a collection of foundational mathematics or early numeracy skills. These early numeracy skills, often referred to as "components" of number sense,

are easy to assess and include skills such as quantity discrimination, one-to-one correspondence, and rote counting. Although researchers have generally agreed that the best way to measure the complex construct of number sense is through the assessment of these early numeracy skills or number sense components, there is still a lack of consensus regarding which components actually represent the construct in a meaningful way (Berch, 2005; Chard et al., 2005; Clarke et al., 2008; Gersten & Chard, 1999; Howell & Kemp, 2005, 2006, 2009; Lago & DiPerna, 2010; Lembke & Foegen, 2009; Methe et al., 2011; Politylo et al., 2011). In fact, a recent, comprehensive review of the number sense literature found forty studies containing thirty-four different proposed components of number sense (Politylo et al., 2011). These components ranged from the ability to estimate and count to skills involving an understanding of equivalency and sequencing. The number sense component that appeared most frequently in the literature was quantity discrimination, or the ability to compare magnitudes (this component appeared in thirty-five of the forty studies reviewed; Politylo et al., 2011). Interestingly, quantity discrimination is often assessed through asking an individual which of two sets is larger; this skill is consistent with the original definition of number sense detailed by Dantzig (1946) and is the skill most often used by cognitive and developmental psychologists to assess number sense. Other frequently-cited components of number sense included estimation, rote counting, simple computation, and number identification. Table 1 includes a small sample of studies from the number sense literature and shows the wide range of components used in operationalizing the construct.

Table 1: Range of components used in operationalizing number sense.

Study	Components of Number Sense
Aunio, Niemivirta, et al. (2006)	Quantity discrimination, classification, one-to-one correspondence, seriation, rote counting, structured counting, resultative counting, applied computation
Chard, Clarke, Baker, Otterstedt, Braun, & Katz (2005)	Rote counting, counting on, skip counting, number identification, number writing, quantity discrimination, missing number
Clarke & Shinn (2004b)	Rote counting, number identification, quantity discrimination, missing number
Howell & Kemp (2005)	Rote counting, counting on, quantity discrimination, number recognition, sequencing, matching numerosity, cardinality, subitizing, one-to-one correspondence, comparing spoken numbers
Jordan, Glutting, & Ramineni (2008)	Enumeration, rote counting, counting principles, number recognition, number knowledge (which number comes before/after), quantity discrimination, nonverbal calculation, simple computation, applied computation
Malofeeva, Day, Saco, & Ciancio (2004)	Rote counting, number identification, one-to-one correspondence, ordinality, estimation, quantity discrimination, part-whole relationships, simple computation
National Mathematics Advisory Panel (2008)	Subitizing, rote counting, estimation, simple computation, quantity discrimination; "advanced type of number sense" includes understanding of place value, composition and decomposition of whole numbers, number properties (e.g., associative), application of principles when problem solving
Yang, Hsu, & Huang (2004)	Quantity discrimination, estimation/use of benchmarks, assessing reasonableness of answer, understanding effect of operations on numbers

Given the wide range of components used to operationalize number sense, Howell and Kemp's (2005, 2006, 2009) line of research attempted to establish a consensus in the educational community regarding which components actually comprise the construct. Using a modified Delphi procedure, which is an anonymous survey method, Howell and Kemp (2005) disseminated a series of questionnaires to twelve Australian educational

researchers in an effort to garner their expert opinion on number sense. Participants were given a list of twenty-five number sense skills and were asked to rate how strongly they agreed that a given component was an essential measure of number sense at the time of school entry. A second round of questionnaires asked which skills were essential components of number sense after the first year of schooling; this questionnaire included the original twenty-five skills, plus eleven additional, advanced skills. At the time of school entry, the highest-ranked skills included one-to-one correspondence, matching numerosity, and rote counting. After one year of schooling, respondents agreed that sequencing numerals, number recognition, and making equivalent groups – among other skills – were essential components of number sense. Perhaps not surprisingly, however, respondents only reached a full consensus on ten of the thirty-six proposed components. For the remaining components, some respondents agreed they were essential parts of the number sense construct and others disagreed.

When Howell and Kemp (2006, 2009) expanded their study and distributed their survey to international researchers, similar results were found. Although international respondents agreed on a handful of components (e.g., cardinality, quantity discrimination, matching numerosity), they did not reach a consensus on the majority of the skills proposed. Although Howell and Kemp were unsuccessful in establishing a consensus within the educational community regarding the true components of number sense, their research did serve to highlight the ongoing debate over the construct as well as the attention that is now being paid to early numeracy skills.

Survey methods like the ones used by Howell and Kemp are just some of the ways researchers have attempted to make sense of the many components of number

sense. Other researchers have taken a factor analytic approach in their quest to better understand the construct. Jordan, Kaplan, Oláh, and Locuniak (2006), for example, conducted factor analyses in an attempt to (1) identify which components fit into the overall construct of number sense and (2) better understand the relationships between the components. Jordan et al. (2006) began by assessing the performance of over 400 kindergarten students on a variety of number sense tasks over four time points during the school year: September, November, February, and April. Skills evaluated included various types of counting (e.g., rote counting, counting on, determining if another's counting was correct or incorrect), knowledge of numbers (e.g., what number comes before 6?), patterns, number recognition, estimation, simple computation problems, and applied computation problems. Following data collection, the researchers conducted an exploratory factor analysis. After reviewing the results, Jordan et al. (2006) proposed a two-dimensional model of number sense. This two-dimensional model emerged across all four time points, both at the beginning and end of the kindergarten year. "Basic number skills" fell in the first dimension of the model and included skills such as counting, number recognition, and estimation. The second dimension of the model consisted of more advanced skills such as simple and applied computation problems. Jordan and her colleagues felt that these two dimensions represented the lower and higher order number sense skills originally proposed by Berch (2005); lower order skills may be related to an innate sense of number and quantity, while higher order skills are often learned after traditional mathematics instruction and help to contribute to a more developed number sense (Berch, 2005; Jordan et al., 2006).

Lago and DiPerna (2010) carried out a similar study in which they tested over 200 kindergarteners on a variety of number sense skills (e.g., estimation, counting, measurement concepts) as well as on rapid automatic naming (RAN) skills. Testing took place in the spring, after students had received several months of formal instruction in kindergarten mathematics. Results from their exploratory factor analysis revealed a two-factor model. Quantity discrimination, counting aloud, number identification, measurement concepts, and nonverbal calculation loaded on the first factor, while the RAN activities loaded on the second factor. These two factors, albeit separate, were moderately correlated ($r = -0.60$; students who performed well on the number sense tasks took less time naming stimuli on the RAN activities). The authors acknowledged that while rapid naming and number sense skills appear to be separate constructs, the two might fall under a higher order factor related to mathematics (Lago & DiPerna, 2010). Interestingly, the skills of estimation and counting objects did not load on either of the two factors found. The researchers hypothesized that estimation may represent a higher order skill, which is why it did not fit with the more basic number skills of the first factor. Perhaps if more skills were tested, a third factor would have emerged. As for the skill of counting objects, the unusual set-up of the counting objects task may have affected student scores, thus partially explaining why this skill did not load on either factor (Lago & DiPerna, 2010).

While survey methods and factor analyses have certainly led the educational community to a better understanding of number sense, the community remains unsure regarding how exactly to define the construct. Despite this ongoing disagreement, the majority of researchers and practitioners do agree that the best way to operationalize the

faculty is through the assessment of a variety of number sense components or early numeracy skills. As a result, the number sense assessments currently used in research and practice are all measures of different early numeracy skills, or number sense components. As will be described in the following section, some of these measures assess a great variety of early numeracy skills, while others only assess three or four.

Number Sense Measures

A review of the literature reveals that there are over fifteen measures of number sense currently used in research and practice. Many of these measures are explicitly described as tests of number sense, while others are labeled as assessments of early numeracy skills. Regardless of the language used, all of the measures do assess some combination of number sense components previously described. The following selection of measures will be reviewed below, as they provide an excellent example of the range of number sense components found in various assessments: Clements' (1984) measures of numerical abilities, the Number Knowledge Test (Okamoto & Case, 1996), the Number Sense Test (Reys et al., 1999), the Number Sense Test (Malofeeva et al., 2004), the Early Numeracy Test (Van Luit & Van de Rijt, 2005), the Number Sense Brief Screener (Jordan et al., 2008), and the Test of Early Numeracy (Clarke & Shinn, 2004b).

Clements' Measures of Numerical Abilities

Clements (1984) was one of the earliest researchers to create a battery of tasks measuring the numerical competence of preschoolers. Although not specifically named a test of number sense, Clements' collection of tasks includes many of the skills and components thought to operationalize the number sense construct. The battery, which

was used as a pre- and post-test measure in a study investigating the effects of two mathematics instruction techniques, consists of ten subtests and fifty-nine items.

The first subtest of Clements' battery is called Rational Counting and simply requires children to count various sets of objects. The next subtest, Choosing More, is a classic quantity discrimination task in which children select the greater of two verbally presented numbers (e.g., which is greater, 5 or 9?) and then identify the set with the greater number of objects. Just After, Just Before, and Between is the third task which asks participants to choose the set of objects that includes one more than the presented set and then one less than the presented set. Children then had to identify the set that fit in between two presented sets. In the Counting On and Counting Back task, participants must determine the correct number of objects in a set after objects are added to and taken away from the set. The fifth task is a one-to-one correspondence activity in which children create a row of seven blocks. For the Identity Conservation subtest, participants are required to judge the number of blocks in a set after they have been scattered. Equivalence Conservation involves judging whether two sets of the same number of objects are still equal after they had been scattered or moved. The final two subtests require children to solve both verbal and concrete word problems.

Although no evidence exists for the validity of these tasks, Clements (1984) did report on the reliability of the measures. Internal consistency was computed using coefficient alpha and was quite high; reliability coefficients ranged from 0.95 for the pre-test to 0.97 for the post-test. The wide variety of number sense components assessed by Clements' ten tasks is impressive, however the large number of items and length of time necessary to administer all ten tasks makes Clements' measures difficult to use as a

screening tool. Nevertheless, Clements' number tasks have served an important role: they were one of the first collections of assessment activities related to the construct of number sense and they have also influenced the development of other, more recently-created measures.

Number Knowledge Test

Developed by Okamoto and Case (1996), the Number Knowledge Test (NKT) assesses a range of skills and is designed to measure a child's understanding of the whole number system. According to Gersten et al. (2005), the NKT is one of the most comprehensive number sense measures currently available, although the test does not appear to be used in practice very frequently. The NKT is based on the idea that children progress through four developmental levels as they begin to understand number, quantity, and numerical operations. These developmental levels are as follows (Okamoto & Case, 1996):

- Level 1, Predimensional: Children are expected to count by rote and to quantify globally but not to connect number and quantity.
- Level 2, Unidimensional: Children are expected to have constructed a mental counting line that integrates their understanding of numbers and quantities.
- Level 3, Bidimensional: Children are expected to be able to work simultaneously with two mental counting lines. This means that they can keep track of "ones" and "tens" while adding or subtracting and understand the relation between them. It also means that they can use one counting line to compute the distance between two points on another counting line, thus constructing the notion of a mathematical "difference."
- Level 4, Integrated Bidimensional: Children can extend their understanding of tens and ones to the full number system. They can also integrate "carrying" or "borrowing" with their mental addition and subtraction and can understand the way in which one difference and another can be related. (pp. 226-227)

The NKT itself is organized into four separate sections that correspond with the developmental levels outlined above. Students who answer more than half of the problems correctly on the first level are permitted to move onto the next level, and so on.

Skills assessed on the predimensional level include counting, simple calculation, and quantity discrimination. The next level assesses similar skills, such as more complex computation problems, additional quantity discrimination items, naming which number is bigger or smaller, and counting backwards. The third level of the NKT involves computation and quantity discrimination with two-digit numbers. This level also asks children to identify, for example, how many numbers there are between two and six. The fourth and final level builds on the third level in asking similar but more complex problems. Children are also asked questions such as "What number comes 10 numbers after 99?" and "What is the smallest 5-digit number?"

Baker, Gersten, Katz, Chard, and Clarke (2002) investigated the predictive validity of the NKT with later performance on the Stanford Achievement Test, Ninth Edition (SAT-9). Researchers found that spring of kindergarten performance on the NKT significantly predicted SAT-9 scores in the spring of first grade. Specifically, NKT performance in kindergarten predicted the total score on the SAT-9 ($r = 0.73$), the score on the Problem Solving subtest of the SAT-9 ($r = 0.64$), and the score on the Procedures subtest ($r = 0.69$). Unfortunately, additional studies exploring the technical adequacy of the NKT have not been conducted.

Number Sense Test (A)

In an effort to explore and compare the number sense of students from Australia, Sweden, Taiwan, and the United States, Reys et al. (1999) created the Number Sense Test (NST). The test is based on six different components of number sense. Researchers described the first component as an understanding of the meaning and size of numbers. Reys et al. (1999) measured this component through quantity discrimination items, such

as “How does 2/5 compare in size to 1/2?” (p. 62). The second component, understanding the use of equivalent representations of numbers, is assessed via questions that ask students to show different ways a number, such as a fraction, can be represented. The next component of number sense measured by the NST is the understanding of the meaning and effect of operations; these items ask students to explain, for example, why a number divided by a decimal results in a larger number. The fourth component’s items tap into a student’s understanding of equivalent expressions, while the fifth component’s items look at a student’s ability to count and compute flexibly. Finally, the sixth number sense component involves the use of measurement benchmarks in estimation. Items assessing this skill ask students to estimate with the help of a benchmark or anchor.

The NST was created for use with both national and international students ranging in age from eight to fourteen. The number of test items vary based on which age group is being tested (variation is from thirty to forty-five items), and test reliability was reported to range from 0.72 to 0.85. Evidence supporting the validity of the NST has not been reported. While items on the NST cover a wide variety of number sense components, these items are clearly more suitable for the assessment of number sense in older children and youth. As a result, the use of the NST as a screening measure for early intervention purposes would likely not be appropriate. The NST does, however, provide an excellent example of a number sense assessment that exists for more advanced students.

Number Sense Test (B)

Five years after Reys et al. (1999) developed their Number Sense Test, Malofeeva et al. (2004) created a measure of the same name. After analyzing various definitions of number sense and previously created measures, Malofeeva et al. (2004) developed a

Number Sense Test designed to measure the early numeracy skills of preschoolers, ages three through five. The test consists of six scales: Counting, Number Identification, Number-Object Correspondence, Ordinality, Comparison, and Addition-Subtraction. The Counting scale includes tasks requiring children to count forwards and backwards, identify mistakes in others' counting, and name the number that comes before and after a given number. The Number Identification scale asks children to both verbally and non-verbally identify various numbers. On the Number-Object Correspondence scale, participants are given several tasks that "measure their ability to assign a unique number word to each counted object" (Malofeeva et al., 2004, p. 652). For the Ordinality scale, children are required to name the position of an object as first, second, third, and so on. The Comparison scale consists of various quantity discrimination tasks (e.g., the child is shown two sets of objects and is asked to name which pile is bigger), and the Addition-Subtraction scale requires the student to complete very simple, concrete computation problems.

Researchers reported that the internal consistency for the Number Sense Test as a whole was 0.98, while coefficient alpha for the separate scales ranged from 0.88 to 0.95. Discriminant validity was also demonstrated, as the six scales were highly correlated with themselves but were not as highly correlated with other measures used in the study (e.g., a vocabulary measure). Malofeeva et al. (2004) acknowledged that in addition to being normed, the technical adequacy of the Number Sense Test needs to be further investigated. Although the Number Sense Test does not currently appear to be used in practice, it does show a great deal of promise, as it is appropriate for use with children as young as three years old and it also surveys a breadth of number sense components.

Consequently, the Number Sense Test could be a valuable tool for identifying students who are struggling in mathematics very early on in their education.

Early Numeracy Test

Originally developed in by Dutch researchers and later translated into Finnish, Chinese, Spanish, and English, the Early Numeracy Test (ENT; Van Luit & Van de Rijt, 2005) is widely used among international researchers of mathematics education. The test, which is designed to assess young children's number sense and early mathematical competence, consists of eight different components: (1) concepts of comparison; (2) classification; (3) correspondence; (4) seriation; (5) using counting words; (6) structured counting; (7) resultative counting; and (8) general knowledge of numbers (Van de Rijt, Van Luit, & Pennings, 1999; Van Luit & Van de Rijt, 2005). The concepts of comparison component assesses the understanding of concepts like most, least, higher, and lower. The classification activities ask the child to group objects based on similarities and differences. Correspondence tasks require the child to utilize one-to-one correspondence skills. The seriation component involves ranking objects or pictures from high to low, more to less, or thick to thin. On the using counting words task, the child must count forwards and backwards as well as count on from a given number. Structured counting tasks involve counting objects and also subitizing, while resultative counting activities ask the child to count objects such as blocks without pointing to or moving them. Finally, the general understanding of numbers component requires children to apply mathematics skills to solve real-life problems depicted in drawings (e.g., You have nine marbles. You lose three marbles. How many marbles do you have left?).

Some evidence of technical adequacy exists for the Dutch and Finnish versions of the ENT. Analysis of the original Dutch version of the ENT revealed that the test is internally consistent, as coefficient alphas for each test component ranged from 0.91 to 0.94 (Van de Rijt et al., 1999). Internal consistency for the Finnish version of the test was reported to be 0.90 (Aunio, Hautamäki, Heiskari, & Van Luit, 2006), while internal consistency for the English version of the test was reported as 0.83 (Van de Rijt et al., 2003). Van de Rijt et al. (1999) stated that because no similar early mathematics assessments exist in the Netherlands, information on the validity of the Dutch version of the ENT was not available. The original test was, however, developed using expert judgment, factor analysis, and item response theory (Van de Rijt et al., 1999). Aunio and Niemivirta (2010) demonstrated preliminary evidence for the predictive validity of the Finnish version of the ENT. Kindergarten students' scores on the ENT explained approximately half of the variance in the mathematics performance of those same students in first grade. Aside from one report of internal consistency from Van de Rijt and her colleagues, no additional research has been published on the reliability and validity of the English version of the ENT.

Number Sense Brief Screener

The Number Sense Brief Screener (NSB) was developed over the course of several years, as researchers sought to create a measure that would identify mathematics difficulties in children in kindergarten and first grade (Jordan et al., 2006; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Jordan et al., 2008). When determining what number sense components to assess as part of the NSB, Jordan and her colleagues (2006) focused on skills that were not only validated by research but also skills that appeared in

elementary school math curricula. Ultimately, items selected for inclusion on the NSB assessed the following six components: counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations (Jordan et al., 2008; Jordan, Glutting, & Ramineni, 2010). The counting knowledge and principles items asks students to rote count, count objects, and identify whether another person's counting strategies are correct or incorrect. Number recognition requires children to identify a variety of two- and three-digit numbers. Number comparisons involves quantity discrimination tasks in which students must determine which number is larger or smaller. The nonverbal calculation tasks ask students to solve various computation problems using colored chips. Finally, story problems and number combinations require students to solve a number of simple computation problems (e.g., "how much is two and one?") (Jordan et al., 2008; Jordan, Glutting, & Ramineni, 2010).

Research regarding the psychometric properties of the NSB has been relatively extensive as well as promising. Jordan, Glutting, Ramineni, and Watkins (2010) administered the NSB to kindergarteners in September, November, February, and April and noted that the test-retest reliability of the measure ranged from 0.78 to 0.86. Test-retest reliability in first grade ranged from 0.80 to 0.84 (Jordan, Glutting, & Ramineni, 2010; Jordan, Glutting, Ramineni, & Watkins, 2010). Analysis of the assessment's internal consistency revealed a coefficient alpha of 0.84 (Jordan et al., 2008). In terms of validity, this measure has been shown to predict mathematics achievement in both first and third grades. Performance on the NSB at the beginning of first grade was predictive of later performance on the mathematics subtests of the Woodcock-Johnson III ($r = 0.72$

at end of first grade; r = 0.70 at end of third grade) (Jordan, Glutting, & Ramineni, 2010). In addition, the NSB was shown to predict mathematics performance above and beyond that of other factors such as age, vocabulary, perceptual reasoning skills, and working memory in both first and third grades (Jordan, Glutting, & Ramineni, 2010). Jordan, Glutting, Ramineni, and Watkins (2010) also demonstrated that kindergarten performance on the NSB was predictive of third grade performance on a high-stakes state test, while Jordan et al. (2008) showed that kindergarten performance on the NSB was predictive of third grade performance on the Math Achievement subtest of the Woodcock-Johnson III (r = 0.65). Divergent validity has been demonstrated, as well, as the NSB was poorly correlated with measures of reading achievement such as the Test of Word Reading Efficiency (Jordan et al., 2008).

Test of Early Numeracy

Of all the number sense assessments reviewed in this section, the Test of Early Numeracy (TEN; Clarke & Shinn, 2004b) is perhaps the measure most widely used in practice in the United States. The TEN, which was originally a collection of tasks known as the Early Mathematics Curriculum-Based Measures, was created in response to the field of education's recent emphasis on the importance of number sense development in young children (Clarke & Shinn, 2004a). Although curriculum-based measures in mathematics already existed (e.g., M-CBM), researchers noted that these measures were not entirely useful in early identification; by the time M-CBM could be used, struggling students were often already in mid to late first grade. Thus, the TEN was also developed to provide educators a tool for the early identification of mathematics difficulties in young children (Clarke & Shinn, 2004a).

The TEN is a set of curriculum-based measures designed to assess the number sense skills of students in kindergarten and first grade. These measures consist of four brief subtests: Oral Counting, Number Identification, Quantity Discrimination, and Missing Number (Clarke & Shinn, 2004b). Oral Counting requires students to correctly count up from zero and name as many digits as they can in one minute. In Number Identification, students are given a sheet of paper with numbers ranging from one to ten and are asked to name as many numbers as possible in one minute. Quantity Discrimination requires students to choose which of two presented numbers is bigger. Finally, in Missing Number, students must orally identify the missing number from a sequence of three digits; for example, students see "7 __ 9" and must be able to identify that the missing number is eight (Clarke & Shinn, 2004b).

A great deal of evidence exists supporting the technical adequacy of the TEN (Clarke & Shinn, 2004a, 2004b). Thirteen-week test-retest reliability was reported to range from 0.79 to 0.85 across subtests. Interrater reliability for each of the subtests was very high and ranged from 0.98 to 0.99. Clarke and Shinn (2004a, 2004b) also reported that alternate form reliability ranged from 0.78 to 0.93. Concurrent validity has been demonstrated for the TEN, as well; the subtests correlated well with other measures of early mathematical skills such as the Number Knowledge Test ($r = 0.70$ to 0.80, fall of first grade administration) and the Woodcock-Johnson III's Applied Problems subtest ($r = 0.64$ to 0.71, fall of first grade administration). First graders' performance on the TEN in the fall was also moderately predictive of spring performance on M-CBM ($r = 0.56$ to 0.70) and on the Applied Problems subtest of the Woodcock-Johnson III ($r = 0.72$ to 0.79) (Clarke & Shinn, 2004a, 2004b). It is important to note that the psychometric

properties outlined above are based on the administration of the TEN to first graders. A small number of studies have investigated the reliability and validity of TEN probes administered in kindergarten. Chard et al. (2005), for example, found preliminary evidence for the concurrent and predictive validity of the TEN kindergarten probes with the Number Knowledge Test (r = 0.50 to 0.69), while Martinez, Missall, Graney, Aricak, and Clarke (2009) demonstrated evidence for the predictive validity of the kindergarten probes with the Stanford 10 Achievement Test (r = 0.31 to 0.46). In addition, Baglici, Codding, and Tryon (2010) reported that the alternate-form reliability of the winter kindergarten probes ranged from 0.84 to 0.91. These researchers also discovered that scores obtained on the Missing Number subtest in the winter of kindergarten significantly and uniquely predicted M-CBM scores in the spring of first grade (r = 0.46; Baglici et al., 2010).

Importance of Assessment, Early Identification, and Predictive Validity

While measures of number sense clearly differ in terms of the breadth and depth of skills assessed, all of the measures do share a common purpose: to identify children who are experiencing – or are at risk for experiencing – difficulties in mathematics. The importance of assessing a child's number sense for the purpose of early identification simply cannot be overstated. Both the National Council of Teachers of Mathematics and the National Mathematics Advisory Panel have highlighted number sense as a vital prerequisite to later success in mathematics (NCTM, 2000; NMAP, 2008). A well-developed number sense allows students to understand number facts and algorithms more quickly, use novel problem-solving strategies, recognize errors, and ultimately perform mathematical computations with greater ease. Those students who enter school with a

poorly developed sense of number are at a greater risk of not only struggling in mathematics during the elementary years, but also experiencing failure in mathematics later in life (NMAP, 2008; National Research Council, 2001). Early identification through assessment is, of course, the first step in preventing this failure. As Jordan and her colleagues noted (2006), "if children's learning needs can be identified early on, we may be able to design interventions that prevent failure in math" (p. 154).

Early identification of students with a weak sense of number would not be possible without psychometrically sound measures designed to assess the construct. As this review outlined in the previous section, several assessments of number sense currently exist, although the amount of evidence demonstrating the reliability and validity of their uses for young children varies. Despite this variability, many measures do possess what is arguably one of the more important psychometric properties of a screening measure: predictive validity. Predictive validity is demonstrated when performance on one measure significantly predicts future performance on one or more criterion measures. Of the assessments previously reviewed, researchers have shown varying degrees of predictive validity in the Number Knowledge Test (NKT), Early Numeracy Test (ENT), Number Sense Brief Screener (NSB), and the Test of Early Numeracy (TEN). Kindergarten performance on the NSB, for example, has been shown to predict mathematics achievement in third grade (Jordan, Glutting, Ramineni, and Watkins, 2010). Predictive validity has also been demonstrated for the Finnish version of the ENT, as kindergarteners' scores on the ENT explained 47% of the variance in their classroom math grades at the end of first grade (Aunio & Niemivirta, 2010). Studies investigating the psychometric properties of the TEN have found evidence for the

predictive validity of this assessment, as well. Kindergarten performance on the TEN has been significantly correlated with later performance on the NKT, Stanford 10 Achievement Test, and M-CBM probes (Baglici et al., 2010; Chard et al., 2005; Martinez et al., 2009).

Why is predictive validity so important? In the case of number sense assessment, predictive validity serves to assure researchers and educators that the assessment in question actually measures skills that contribute or are related to future mathematics development and success. As a result, if a measure possesses predictive validity, researchers and educators can feel confident in using that assessment as a screening tool to identify children who may have difficulties in mathematics later in life. In addition, as Gersten, Clarke, Haymond, and Jordan (2011) note, "assessments that show evidence of predictive validity can inform instructional decision-making" (p. 3). Alternatively, if a number sense assessment failed to demonstrate predictive validity, then it would not be measuring skills that contributed to or were related to later success in mathematics. A measure lacking in predictive validity would most likely not be very useful as a screening instrument for early identification purposes.

Summary

As this review has shown, a wealth of research has been conducted on number sense since Tobias Dantzig first coined the term in the early twentieth century. Over the past several decades, researchers have demonstrated that a wide range of animals – from birds to dolphins to chimps – possess an innate sense of number; they have shown that this same innate number sense also exists in human infants; and they have uncovered links between children's early number sense abilities and their later success in higher-

level mathematics. Despite this abundance of research, the construct of number sense remains a somewhat nebulous one. While developmental and cognitive psychologists tend to define number sense the same way it was originally described by Dantzig (i.e., the ability to identify a difference in size between two small groups and to recognize when an element has been removed or added from a group), the definitions of number sense used in the field of education are much more complex, vague, and varied. Some educational researchers define number sense more broadly and conceptually; for example, Howden (1989) described number sense as having a "good intuition about numbers and their relationships" (p. 11). Many others, however, have defined the construct operationally. Operational definitions of number sense typically consist of a collection of early numeracy skills or "number sense components" such rote counting, quantity discrimination, number identification, and one-to-one correspondence.

Although there remains a lack of consensus regarding exactly how to define number sense, research suggests that a well-developed number sense is a vital prerequisite to later success in mathematics (NCTM, 2000; NMAP; 2008). Without this well-developed sense of number, students are at a greater risk of struggling and failing in mathematics (NMAP, 2008; National Research Council, 2001). As a result, researchers have developed various number sense assessments in recent years in an effort to screen for and identify children who may be at risk for later failure in mathematics. These measures include, but are not limited to, Clements' measures of numerical abilities, the Number Knowledge Test, the Number Sense Test, the Early Numeracy Test, the Number Sense Brief Screener, and the Test of Early Numeracy. While Gersten et al. (2005) states that "research on valid early screening measures of subsequent mathematics proficiency

is in its infancy" (p. 293), the need for these types of measures is clear. If researchers can determine which number sense assessment best predicts later success in mathematics, then practitioners would have a powerful tool for screening and identifying young children who are at risk for experiencing difficulties in mathematics. Ideally, this early identification would then lead to early intervention, which may in turn prevent later failure in mathematics.

This study was conducted in an effort to investigate the predictive validity of three of the more widely used measures of number sense: the Early Numeracy Test (ENT), the Number Sense Brief Screener (NSB), and the Test of Early Numeracy (TEN). Although preliminary evidence for the predictive validity of these three measures currently exists, no studies have compared the predictive qualities of these measures directly. As a result, this study sought to identify which measure or combination of measures best predicts later mathematics achievement. If one or more measures can be identified as the best predictor(s) of later mathematics achievement, then educators would be equipped with an important and valid tool to be used in the early identification of mathematics difficulties. In addition, this study was designed to add to the extant literature on the ENT, NSB, and TEN. While a small body of research currently exists on each measure, few studies have replicated this work. Furthermore, while the ENT is popular abroad, there has been no research conducted on the measure in the United States.

CHAPTER 3

METHODOLOGY

Participants and Setting

A Priori Analyses

This study was conducted during the 2012-2013 school year in a suburban school district in Western Massachusetts. The school district's superintendent and special education director gave the researcher permission to conduct the study in two elementary schools, each enrolling approximately 75 kindergarteners for a total of 150 possible participants.

In order to determine if a sample size of 150 participants would be sufficient to conduct significance tests, a priori power analyses were conducted for each statistical method used in the study. The alpha level was held constant at 0.05 across all analyses. The statistical software R (R Development Core Team, 2012) was used to conduct the a priori analysis for a simple linear regression. Results revealed that given a sample size of 150 participants, there would be adequate power (0.98) for detecting a moderate effect size. The power analysis software G*Power (Faul, Erdfelder, Lang, & Buchner, 2009) was used to conduct the a priori analysis for comparing dependent correlations with a common index. This analysis showed that given a sample size of 150 participants, there would be sufficient power (0.79) for detecting a small to moderate difference in correlations. To evaluate whether there would be adequate power to conduct multiple regression analyses given a sample size of 150, a final a priori analysis was conducted using G*Power. Results from the analysis indicated that there would be sufficient power (0.97) to detect a small to moderate effect size.

Recruitment

Three days before the school year began, each participating school held kindergarten orientation sessions for parents and students. During these orientation sessions, the researcher and school principal presented the research study to parents, answered questions, and distributed consent forms further describing the study (Appendix A). School staff translated consent forms for parents who spoke a language other than English. Parents were given the option to immediately return the consent form at the orientation session or return the form to school with their child by the first day of school. Parents who did not return the consent form were called by the researcher and reminded to submit the form as soon as possible. All parents were assured that participation in the study was voluntary, and that they could withdraw permission for their child's participation at any time. Of the 148 kindergarteners enrolled in the two elementary schools during the 2012-2013 academic year, 112 were given permission to participate.

Before testing began, verbal assent was obtained from each participant. Kindergartners were given a brief overview of each testing activity and were asked if they wanted to participate. After listening to a description of the first assessment, one student chose not to participate and was consequently excluded from the sample.

Participant and School Characteristics

While consent was obtained for 112 kindergarteners, seven students moved out of the district mid-year, two students were non-English speakers, and one student chose not to participate. Thus, participants were ultimately 102 kindergarteners from both general and special education. Non-English speaking students were excluded from the sample as this study aimed to assess mathematical skills, not understanding of the English language.

All assessments used required an understanding of the English language and the majority of assessment items involved a verbal response in English.

As noted, this study took place in two mid-sized, suburban elementary schools in Western Massachusetts. The first elementary school serves approximately 350 students in kindergarten through grade four. During the 2012-2013 academic year, 55% of the student body was female and 45% was male. In terms of ethnicity, 86% of the students were white, 5% Asian, 4% multiracial, 3% Hispanic, and 2% African American. Approximately 26% of the students in the school were eligible for free or reduced-price lunch and 7% received special education services (Massachusetts Department of Elementary and Secondary Education [DESE], 2013). The second elementary school serves approximately 400 students in kindergarten through grade four. During the 2012-2013 academic year, 51% of the student body was female and 49% was male. In terms of ethnicity, 88% of the students were white, 5% Hispanic, 3% multiracial, 2% Asian, and 2% African American. Approximately 36% of the students in the school were eligible for free or reduced-price lunch and 13% received special education services (Massachusetts DESE, 2013).

Both participating schools operate under a response-to-intervention (RTI) framework and utilize a three-tired service delivery model. Students in the first tier (i.e., all students) receive high-quality instruction from evidence-based curricula and their progress is measured three times per year through universal benchmarking. At the time of this study, the participating kindergarten classrooms were using Scott Foresman-Addison Wesley's *enVisionMATH* curriculum. Kindergarteners' progress in mathematics was measured in the fall, winter, and spring using the Test of Early Numeracy (TEN)

curriculum-based measures. If students were not making effective progress in the curriculum, they received small group mathematics interventions three times per week for one half hour, administered by the district's primary preventionists. This represents the second tier of services. Of this study's 102 participants, nine received additional support in mathematics from the primary preventionists at some point during the school year. Students who continued to struggle in the curriculum despite primary prevention support were then referred for a special education evaluation. If a student was found eligible, special education services in mathematics – the third and highest tier of services – were provided. Four of the 102 participants in this study received special education services in mathematics during the 2012-2013 academic year.

Independent Variables

Although the Test of Early Numeracy, Number Sense Brief Screener, and Early Numeracy Test were described in the previous chapter, they will again be briefly reviewed here. The following descriptions revisit each tool's psychometric properties and also provide more specific information regarding the administration procedures for each measure.

Test of Early Numeracy

The first number sense measure administered to students was the Test of Early Numeracy (TEN). The TEN is a set of individually administered curriculum-based measures designed to assess the early numeracy skills of students in kindergarten and first grade. These measures consist of four subtests that each take one minute to complete: Oral Counting, Number Identification, Quantity Discrimination, and Missing Number (Clarke & Shinn, 2004b). Oral Counting requires the student to correctly count

up from zero and name as many digits as he or she can in one minute. In Number Identification, students are given a sheet of paper with numbers ranging from 1 to 10, randomly arranged in an 8 by 7 array, and are asked to name as many numbers as possible in one minute. Quantity Discrimination requires students to choose which of two presented numbers is bigger. Finally, in Missing Number, students must orally identify the missing number from a sequence of three digits; for example, students see "7 __ 9" and must be able to identify that the missing number is eight (Clarke & Shinn, 2004b). It is important to note that the TEN is the screening assessment typically used by the participating schools for universal benchmarking and progress monitoring.

Clarke and Shinn (2004a, 2004b) have provided evidence that the TEN is both reliable and valid for the assessment of the early numeracy skills of first graders. The subtests of the TEN have a thirteen-week test-retest reliability ranging from 0.79 to 0.85. Interrater reliability ranges from 0.98 to 0.99. Clarke and Shinn (2004b) also reported alternate form reliability ranging from 0.78 to 0.93. Concurrent validity has been demonstrated for the TEN, as well; the subtests correlate well with other measures of early mathematical skills such as the Number Knowledge Test ($r = 0.70$ to 0.80, fall of first grade administration) and the Woodcock-Johnson III's Applied Problems subtest ($r = 0.64$ to 0.71, fall of first grade administration) (Clarke & Shinn, 2004b). First graders' performance on the TEN in the fall was also moderately predictive of spring performance on mathematics curriculum-based measures (M-CBM; $r = 0.56$ to 0.70) and on the Applied Problems subtest of the Woodcock-Johnson III ($r = 0.72$ to 0.79) (Clarke & Shinn, 2004b). A small number of studies have investigated the reliability and validity of TEN probes administered in kindergarten, as well. Chard et al. (2005) found preliminary

evidence for the concurrent and predictive validity of the TEN kindergarten probes with the Number Knowledge Test (r = 0.50 to 0.69), while Martinez et al. (2009) demonstrated evidence for the predictive validity of the kindergarten probes with the Stanford 10 Achievement Test (r = 0.31 to 0.46). In addition, Baglici, Codding, and Tryon (2010) reported that the alternate-form reliability of the winter kindergarten probes ranged from 0.84 to 0.91. These researchers also discovered that scores obtained on the Missing Number subtest in the winter of kindergarten significantly and uniquely predicted M-CBM scores in the spring of first grade (r = 0.46; Baglici et al., 2010).

Number Sense Brief Screener

The next number sense measure administered to participating kindergarteners was the Number Sense Brief Screener (NSB). The NSB is a 33-item, individually administered assessment of number sense intended for use with kindergarteners and first graders (Jordan et al., 2008). The measure includes items on rote counting, one-to-one correspondence, counting principles, number recognition, number comparison, nonverbal calculation, story problems, and simple computations. Each item is marked either correct or incorrect, and no partial credit is given for any item. The NSB is untimed and takes approximately 15 minutes to administer.

Plenty of evidence exists supporting the technical adequacy of the NSB. Jordan, Glutting, Ramineni, and Watkins (2010) administered the NSB to kindergarteners four separate times throughout the course of one year and reported the test-retest reliability of the measure to range from 0.78 to 0.86. Coefficient alpha was reported as 0.84, demonstrating acceptable internal consistency (Jordan et al., 2008). In terms of validity, this measure has been shown to predict mathematics achievement in both first and third

grades. Performance on the NSB at the beginning of first grade was predictive of later performance on the mathematics subtests of the Woodcock-Johnson III ($r = 0.72$ at end of first grade; $r = 0.70$ at end of third grade) (Jordan, Glutting, & Ramineni, 2010). Kindergarten performance on the measure was also predictive of scores on a high-stakes state test administered in third grade (Jordan, Glutting, Ramineni, and Watkins, 2010). In addition, Jordan et al. (2008) showed that kindergarten performance on the NSB was predictive of third grade performance on the Math Achievement subtest of the Woodcock-Johnson III ($r = 0.65$). Divergent validity has been demonstrated for the NSB, as well, as the measure does not correlate well with a test of reading achievement (Jordan et al., 2008).

Early Numeracy Test

The third and final number sense measure given to participants was the Early Numeracy Test (ENT). Originally developed by Dutch researchers and later translated into Finnish, Chinese, Spanish, and English, the ENT is a 40-item, individually administered assessment designed to measure the early mathematical competence of students in preschool through first grade (Van Luit & Van de Rijt, 2005). The assessment measures eight components of early mathematical competence: (1) concepts of comparison; (2) classification; (3) correspondence; (4) seriation; (5) using counting words; (6) structured counting; (7) resultative counting; and (8) general knowledge of numbers. Each item is marked either correct or incorrect, and no partial credit is given for any item. The ENT is untimed and takes approximately 30 minutes to administer. The test is also available in two alternate forms.

Analysis of the original Dutch version of the ENT revealed that the test was internally consistent, with coefficient alphas for each component ranging from 0.91 to 0.94 (Van de Rijt et al., 1999). Internal consistency for the Finnish version of the test was reported to be 0.90 (Aunio, Hautamäki, et al., 2006), while internal consistency for the English version of the test was reported as 0.83 (Van de Rijt et al., 2003). Van de Rijt et al. (1999) stated that because no similar early mathematics assessments exist in the Netherlands, information on the validity of the ENT is not available. The original test was, however, developed using expert judgment, factor analysis, and item response theory (Van de Rijt et al., 1999). Aunio and Niemivirta (2010) demonstrated preliminary evidence for the predictive validity of the Finnish version of the ENT; kindergarteners' scores on the ENT explained approximately half of the variance in the mathematics performance of those same students in first grade. Aside from the one report of internal consistency from Van de Rijt and her colleagues, no additional research has been published on the reliability and validity of the English version of the ENT.

Teacher Rating

In addition to measuring participants' number sense directly, classroom teachers were given a simple rating scale and were asked to rate the number sense of each participant at both the beginning and end of the school year (Appendix B). On the rating scale, teachers were given a definition of number sense from the National Mathematics Advisory Panel (2008) and were then asked to rate the number sense of each participant to the best of their ability on a Likert scale that ranged from 1 to 10. One represented a poorly developed number sense and ten represented a well-developed number sense.

Criterion Measure

Mathematics achievement was measured using the Test of Early Mathematics Ability, Third Edition (TEMA-3; Ginsburg & Baroody, 2003). The TEMA-3 is an individually administered, norm-referenced test intended for use with children ages 3 years 0 months to 8 years 11 months. The assessment has two parallel forms with 72 items each. Items on each form are designed to assess skills in the domains of number, number comparison, numeral literacy, number facts, simple calculation, and general understanding of mathematical concepts. Each item is marked either correct or incorrect, and no partial credit is given for any item. If students receive a score of zero on five consecutive items, the test is discontinued. The TEMA-3 is untimed and takes approximately 30 to 40 minutes to fully administer.

Given the well-established nature of the TEMA-3, the assessment possesses sound psychometric properties. Test-retest and alternate form reliabilities are both above 0.80 (Ginsburg & Baroody, 2003). The test is also internally consistent, as all reported reliability coefficients fall above 0.92. Concurrent validity has been demonstrated by correlating the TEMA-3 with other popular measures of early mathematics abilities. The TEMA-3 correlates well with the Young Children's Achievement Test ($r = 0.91$) and the Math Calculation subtest of the Diagnostic Achievement Battery ($r = 0.83$). Moderately strong correlations are also reported for the Basic Operations subtest of the Key Math assessment ($r = 0.63$) and the Woodcock-Johnson III Tests of Achievement ($r = 0.55$) (Ginsburg & Baroody, 2003).

Procedures

Training of Data Collectors

Prior to the first phase of data collection, six school psychology graduate students were trained in the administration and scoring of the three number sense measures. During each training session, trainees were given a brief overview of the study and were then instructed on how to administer, score, and check the administration integrity of each number sense measure. Trainees then paired off and practiced administering the number sense measures together; that is, one trainee acted as the student and also checked administration integrity while the other administered and scored the number sense assessment. Each trainee practiced administering the assessments until administration integrity exceeded 90% on three consecutive administrations. At the conclusion of each training session, trainees were sent an electronic copy of the assessment instructions, probes, protocols, and integrity checklists for review.

Training on the criterion measure occurred in the spring before the second phase of data collection. During these spring training sessions, five school psychology graduate students and one school counseling graduate student were instructed on how to administer, score, and check the administration integrity of the criterion measure, the TEMA-3. Trainees again paired off and practiced administering the TEMA-3 together; one trainee acted as the student and also checked administration integrity while the other administered and scored the assessment. Each trainee practiced administering the TEMA-3 until administration integrity exceeded 90% on three consecutive administrations. At the conclusion of each training session, trainees were given a copy of the TEMA-3 test protocol, instructions, and integrity checklists for review.

Data Collection

The first phase of data collection began in mid September, during the first full week of kindergarten, and lasted two consecutive weeks (Table 2). During this phase, the TEN, NSB, and ENT were individually administered to each participating kindergarten student over the course of three different testing sessions. Teacher rating scales were also distributed and collected during this phase.

Table 2: Timeline of fall data collection.

	Activity
Week 1	Distribute teacher rating scales Administer TEN Administer NSB
Week 2	Administer ENT Collect teacher rating scales

Fall data collection began with the administration of the TEN. Test administrators followed the standardized instructions for all four of the minute-long TEN subtests and recorded the student's score and number of errors after each administration. Although alternate forms exist for the TEN, the same forms were given to all kindergarteners as part of this study. After the TEN was administered to all participating kindergarteners, data collectors then administered and scored the NSB during a separate, second testing session. Over the course of the second week of fall data collection, the final number sense measure, the ENT, was administered and scored according to the standardized instructions. While two alternate forms are available for the ENT, Form A was administered to all participants in this study.

Fall assessment administration was purposely not counterbalanced in order to ensure that the amount of time between each assessment was the same for each participant. In other words, students who were administered the TEN first were then

administered the NSB first, and so on. Data collectors pulled students for testing alphabetically so that order of testing remained consistent and easy to track.

The second and final phase of data collection began in late May, lasted two consecutive weeks, and ended approximately two weeks before the last day of school (Table 3). During this phase, the criterion measure, the TEMA-3, was individually administered to each participating kindergarten student according to standardized instructions. The same teacher rating scales that were distributed during fall data collection were also distributed and collected during this phase. To again ensure equal amounts of time between each testing session, students who were tested first in the fall were tested first in the spring.

Table 3: Timeline of spring data collection.

	Activity
Week 1	Distribute teacher rating scales Begin administering TEMA-3
Week 2	Finish administering TEMA-3 Collect teacher rating scales

Administration Integrity and Inter-rater Reliability

Administration integrity and inter-rater reliability were checked for 23% of the test administrations. For approximately one out of every five assessment administrations, a secondary data collector was present to (1) ensure administration integrity via an integrity checklist, and (2) simultaneously and independently score the assessment along with the primary data collector. Administration integrity checklists already existed for the TEN (Clarke & Shinn, 2004b) and were created by the researcher for the NSB, ENT, and TEMA-3 (Appendix C). The secondary data collector simply used an extra test protocol to simultaneously and independently score the assessment for inter-rater reliability purposes.

Administration integrity was calculated using the following formula: (Number of steps completed correctly)/(Total number of steps) x 100%. Results indicated that both the fall and spring measures were administered with a high degree of integrity. Administration integrity for the subtests of the TEN ranged from 97% to 99%, while the administration integrity for the NSB and ENT averaged 99%. Integrity of administration for the criterion measure, the TEMA-3, averaged 96%.

Inter-rater reliability was calculated using a standard response-by-response or point-by-point agreement formula where the number of agreements was divided by the number of agreements plus the number of disagreements (i.e., Inter-rater Reliability = Agreements/(Agreements + Disagreements)). Inter-rater reliability for the subtests of the TEN ranged from 0.90 to 0.97, suggesting a high level of agreement between raters across subtests. Inter-rater reliability for the NSB and ENT was 0.98 and .99, respectively. Inter-rater reliability for the TEMA-3 was 0.99, which also indicated a very high level of agreement between raters for this measure.

Data Analytic Plan

Calculating a TEN Composite Score

As previously described, the Test of Early Numeracy (TEN) consists of four separate subtests: Oral Counting, Number Identification, Quantity Discrimination, and Missing Number. While each of these subtests can be administered and scored independently, all four subtests are typically administered together when the TEN is being used for screening purposes. In an effort to determine the predictive validity of the TEN as a whole, as opposed to the predictive validity of each separate subtest, a composite score was calculated for the TEN. Several methods were considered for

creating this composite score (Appendix D); however, all methods produced composite scores that were highly correlated with one another and thus the simplest method was chosen to create the composite score. This method involved summing the scores obtained on each of the four subtests to create the overall TEN composite score. For example, if a student scored a 50 on Oral Counting, 25 on Number Identification, 10 on Quantity Discrimination, and 5 on Missing Number, the student's TEN composite score would be 90.

Preliminary Analyses

Prior to conducting the analyses related to the primary research questions, data were screened using SPSS (Version 22.0) to determine if the underlying assumptions of simple linear and multiple regression analyses would be met. More specifically, the assumptions of normality, linearity, and homoscedasticity were assessed by examining histograms, scatterplots, Q-Q plots, plots of the residuals versus predicted values, and skew and kurtosis statistics. Multicollinearity among independent variables was also assessed through examination of the correlation matrix and variance inflation factors (VIF). Univariate outliers were identified using two methods: (1) visually inspecting boxplots for each univariate distribution and (2) converting each variable's raw scores to *z* scores. *Z* scores beyond ±3.29 were considered outliers ($p < 0.001$; Tabachnick & Fidell, 2001). Bivariate outliers were identified through the calculation of Mahalanobis distances for each multiple regression analysis.

Primary Analyses

Following the preliminary analyses, a number of additional statistical analyses were conducted in order to answer this study's primary research questions. All analyses

were run in SPSS, aside from the analysis that involved comparing dependent correlations. This particular analysis was conducted using R's 'psych' package (R Development Core Team, 2012; Revelle, 2013). One-tailed significance tests were used for all regression analyses.

Simple linear regression analyses were conducted to examine the relationship between kindergarteners' fall performance on each number sense measure and spring performance on the TEMA-3. A simple linear regression was also used to assess the relationship between fall teacher rating of number sense and spring performance on the TEMA-3. To determine which number sense measure best predicted later mathematics achievement, dependent correlations between performance on the number sense measures and the TEMA-3 were compared using the methods described in Steiger (1980). Multiple regression analyses were then run to examine whether or not certain combinations of number sense measures and/or fall teacher rating of number sense predicted mathematics achievement above and beyond that of just one measure (e.g., did the TEN and the NSB predict mathematics achievement above and beyond that of the NSB alone?). Finally, a simple linear regression was used to determine if there was a significant relationship between spring teacher rating of number sense and spring performance on the TEMA-3.

CHAPTER 4

RESULTS

Summary of Purpose

The purpose of this study was to investigate the predictive utility of three measures of number sense: the Test of Early Numeracy (TEN), Number Sense Brief Screener (NSB), and Early Numeracy Test (ENT). It was hypothesized that there would be a positive relationship between fall performance on each number sense measure and spring performance on a mathematics achievement test, the Test of Early Mathematics Ability, Third Edition (TEMA-3). This hypothesis was tested using simple linear regression analyses. It was also hypothesized that the ENT would emerge as the best predictor of number sense, as it is the most comprehensive measure of number sense and assesses the broadest range of early numeracy skills. This hypothesis was tested using the methods described in Steiger (1980) for comparing dependent correlations. In addition to examining the predictive utility of each number sense measure, this study analyzed the predictive validity of teacher rating of student number sense in the fall. It was predicted that there would be a positive relationship between fall teacher rating of number sense and spring performance on the TEMA-3. This prediction was tested using a simple linear regression analysis.

This study also aimed to determine if certain combinations of number sense measures predicted mathematics achievement above that of just one measure. It was hypothesized that all possible combinations of number sense measures would predict mathematics achievement above and beyond that of just one measure. This hypothesis was developed due to the fact that all three measures of number sense, while assessing

the same construct, all contain at least a few unique items that measure different early numeracy skills. It was also hypothesized that performance on a number sense measure combined with teacher rating of number sense would predict mathematics achievement above and beyond that of performance on a number sense measure alone. These predictions were tested through multiple regression analyses.

Finally, this study sought to examine the relationship between a teacher's rating of a kindergartener's number sense in the spring and that same kindergartener's performance on a mathematics achievement test in the spring. It was hypothesized that there would be a positive relationship between these variables; a simple linear regression was used to test this hypothesis.

Screening for Assumptions

Data were screened for normality, linearity, homoscedasticity, outliers, and mulicollinearity in an effort to determine if the underlying assumptions of simple linear and multiple regression analyses would be met. Normality for each univariate distribution was assessed via the visual inspection of histograms and Q-Q plots (Appendices E and F, respectively) as well as analysis of skew and kurtosis statistics (Table 4). Results revealed that each independent variable was approximately normally distributed, aside from a slightly non-normal distribution for teacher rating of number sense in the spring due to a small negative skew (-1.07). The distribution of the dependent variable, the TEMA-3, was somewhat non-normally distributed due to a slight positive skew (1.02) and moderately high kurtosis (3.88). Closer examination of this distribution showed that two outliers were likely the cause of the positive skew and heavy

tails. Despite this skewness and kurtosis, data were considered normal enough to conduct the intended analyses without negative consequences.

Table 4: Descriptive statistics for independent and dependent variables.

Variable	Min.	Max.	Mean	Std. deviation	Skewness[a]	Kurtosis[b]
Independent Variables						
TEN Composite	10	201	97.41	46.44	0.07	-0.78
NSB	8	33	17.67	5.73	0.37	-0.52
ENT	7	36	19.68	6.48	0.33	-0.46
T. Rating (Fall)	2	10	6.91	1.92	-0.63	0.17
T. Rating (Spring)	2	10	7.59	2.06	-1.07	0.68
Dependent Variable						
TEMA-3	17	71	35.84	8.73	1.02	3.88

[a]The standard error of the skewness was 0.24.
[b]The standard error of the kurtosis was 0.47.

Linearity was evaluated through the visual inspection of scatterplots showing the relationship between the dependent variable and each independent variable (Appendix G). Scatterplots revealed linear relationships in all cases. Homoscedasticity was tested by creating a plot of the residuals versus predicted values for each simple linear regression model (Appendix H). Results indicated consistent error variance around the regression line for each model, and thus the assumption of homoscedasticity was satisfied.

Multicollinearity among independent variables was first assessed through examination of the correlation matrix, which showed correlations between independent variables ranging from 0.39 to 0.83 (Table 5).

Table 5: Correlation matrix.

	TEN Composite	NSB	ENT	T. Rating (Fall)	T. Rating (Spring)	TEMA-3
TEN Composite	1.00	0.81	0.83	0.49	0.66	0.73
NSB		1.00	0.78	0.45	0.65	0.78
ENT			1.00	0.39	0.63	0.71
T. Rating (Fall)				1.00	0.58	0.43
T. Rating (Spring)					1.00	0.65
TEMA-3						1.00

Variance inflation factors (VIF) were also examined. These factors did not exceed 4.0 in any of the multiple regression analyses run.

Univariate outliers were identified using two methods: (1) visually inspecting boxplots for each univariate distribution and (2) converting each variable's raw scores to *z*-scores and comparing the *z*-scores to a critical value of ±3.29 ($p < 0.001$; Tabachnick & Fidell, 2001). Analysis of the boxplots and *z*-scores did not indicate any outliers for the TEN Composite, NSB, or ENT distributions. Boxplots of fall and spring teacher rating of number sense identified three and five outliers, respectively. The *z*-scores for these outliers did not, however, exceed ±3.29, and thus these outliers were not removed from the sample. A boxplot of the TEMA-3 distribution identified three outliers, but only one of these three outliers exceeded the critical value of ±3.29. Further examination revealed that the extreme score was simply due to this particular participant performing very well on the assessment and was not the result of an error in data entry or unstandardized administration of the assessment. Consequently, this outlier was not removed from the sample. The calculation of Mahalanobis distances for each multiple regression analysis did not reveal any bivariate outliers.

Research Question One

Is there a relationship between kindergarteners' fall performance on each measure of number sense and spring performance on a mathematics achievement test?

Simple linear regression analyses were conducted to determine if there was a significant relationship between fall performance on each number sense measure and spring performance on the TEMA-3. It was hypothesized that there would be a positive relationship between fall performance on each measure of number sense and spring

performance on the TEMA-3. Analyses revealed positive and strong relationships between fall performance on each number sense measure and spring performance on the TEMA-3, with standardized regression coefficients (R) ranging from 0.71 to 0.78 (Table 6). In addition, results indicated that each regression model was significant ($p < 0.001$), suggesting that the TEN Composite, NSB, and ENT each significantly predict later performance on the TEMA-3. A closer examination of the R^2 values showed that each number sense measure also explained approximately half of the variability in TEMA-3 scores. The NSB explained the most variance, with an R^2 value of 0.61, while the ENT explained the least variance ($R^2 = 0.50$).

Table 6: Simple linear regression analyses between TEMA-3 and each number sense measure ($TEMA\text{-}3_i = \beta_0 + \beta_1(Predictor_i) + \varepsilon_i$).

	Unstandardized Coefficients						95% Confidence Interval for β	
Predictor	β	SE	R	R^2	t	p	Lower	Upper
TEN Composite	0.14	0.01	0.73	0.54	10.80	<0.001	0.11	0.16
NSB	1.19	0.10	0.78	0.61	12.40	<0.001	1.00	1.38
ENT	0.95	0.10	0.71	0.50	9.99	<0.001	0.76	1.14

Research Question Two

Which measure of number sense, administered in the fall of kindergarten, best predicts mathematics achievement in the spring of kindergarten?

To determine which of the three number sense measures best predicts later mathematics achievement, dependent correlations between performance on the number sense measures and the TEMA-3 were compared using the methods described in Steiger (1980). Since the ENT is the most comprehensive measure of number sense and assesses the broadest range of early numeracy skills, it was hypothesized that the ENT would be the best predictor of later performance on the TEMA-3. Results from the analysis showed no significant differences in the correlations between the TEN Composite, NSB,

and ENT and the TEMA-3. In other words, no single number sense measure predicted later mathematics achievement significantly better than the others (Table 7).

Table 7: Comparing dependent correlations.

Comparison	Correlation with TEMA-3 (A)	Correlation with TEMA-3 (B)	t	p
TEN Composite (A) vs. NSB (B)	0.73	0.78	1.17	0.24
TEN Composite (A) vs. ENT (B)	0.73	0.71	0.70	0.49
NSB (A) vs. ENT (B)	0.78	0.71	1.77	0.08

Although no number sense measure emerged as the best predictor of later mathematics achievement in a statistical sense, the NSB has advantages over the other measures. For example, according to anecdotal evidence gathered from data collectors, this measure was the simplest to administer, easiest to score, and appeared to be the most engaging to students. As a result, the NSB was treated as the "best" measure when conducting analyses for research questions four and five.

Research Question Three

Is there a relationship between a teacher's rating of a kindergartener's number sense in the fall and that same kindergartener's performance on a mathematics achievement test in the spring?

A simple linear regression analysis was conducted to determine if there was a relationship between fall teacher rating of number sense and spring performance on the TEMA-3. It was hypothesized that there would be a positive relationship between a teacher's rating of a kindergartener's number sense in the fall and that same kindergartener's spring performance on the TEMA-3. The analysis revealed a positive, moderately strong relationship between fall teacher rating of number sense and later mathematics achievement ($R = 0.43$; Table 8). Additionally, results indicated the

regression model was significant ($p < 0.001$), suggesting that fall teacher rating of number sense significantly predicts later performance on the TEMA-3. The R^2 value, however, showed that fall teacher rating explained only 19% of the variability in TEMA-3 scores.

Table 8: Simple linear regression analyses between TEMA-3 and fall teacher rating of number sense ($TEMA\text{-}3_i = \beta_0 + \beta_1(Fall\ Rating_i) + \varepsilon_i$).

	Unstandardized Coefficients						95% Confidence Interval for β	
Predictor	β	SE	R	R^2	t	p	Lower	Upper
Fall rating	1.97	0.41	0.43	0.19	4.81	<0.001	1.16	2.78

Given that fall teacher rating of number sense predicted performance on the TEMA-3, exploratory analyses were conducted to determine if there was a significant difference in the way fall teacher rating predicted TEMA-3 performance compared to the way the number sense measures predicted TEMA-3 performance. That is, does a teacher's rating of a student's number sense in the fall predict later mathematics achievement better than the TEN Composite, NSB, and/or ENT? Using the methods described in Steiger (1980) for comparing dependent correlations, the correlation between fall teacher rating and the TEMA-3 was compared to the correlations between each number sense measure and the TEMA-3. Results revealed that the TEN Composite, NSB, and ENT all predict TEMA-3 performance significantly better than fall teacher rating of number sense (Table 9).

Table 9: Comparing dependent correlations between the TEMA-3 and fall teacher rating of number sense versus the TEMA-3 and the number sense measures.

Comparison	Correlation with TEMA-3 (A)	Correlation with TEMA-3 (B)	t	p
Fall Teacher Rating (A) vs. TEN Composite (B)	0.43	0.73	4.27	<0.001
Fall Teacher Rating (A) vs. NSB (B)	0.43	0.78	5.04	<0.001
Fall Teacher Rating (A) vs. ENT (B)	0.43	0.71	3.43	0.001

Research Question Four

Is there a combination of number sense measures that predicts mathematics achievement above and beyond that of just one measure?

Multiple regression analyses were conducted to determine if certain combinations of number sense measures predict mathematics achievement above that of just one measure. It was hypothesized that each combination of number sense measures created would predict mathematics achievement above and beyond that of just one measure alone. For these analyses, the proposed method was to place the best predictor of number sense (as determined by the results from research question two) in the model first, followed by different combinations of the other two number sense measures. Although the results from research question two showed that no number sense measure emerged as the best predictor of later mathematics achievement in a statistical sense, it was determined that the NSB could be considered the best measure when comparing the three measures qualitatively. As a result, the NSB was placed in each regression model first when creating combinations of number sense measures.

Results indicated that all three models predicted mathematics achievement above that of just the NSB alone (Table 10).

Table 10: Multiple regression analyses of different combinations of number sense measures.

Predictor	Unstandardized Coefficients β	SE	t	p	R^2/ (R^2 Change)
$TEMA\text{-}3_i = \beta_0 + \beta_1(NSB_i) + \beta_2(TEN\ Composite_i) + \varepsilon_i$					
Constant	15.91	1.74	9.12	<0.001	0.64/ (0.03)
NSB	0.81	0.16	5.24	<0.001	
TEN Composite	0.06	0.02	2.99	0.002	
$TEMA\text{-}3_i = \beta_0 + \beta_1(NSB_i) + \beta_2(ENT_i) + \varepsilon_i$					
Constant	13.53	1.81	7.48	<0.001	0.63/ (0.02)
NSB	0.89	0.15	5.92	<0.001	
ENT	0.34	0.13	2.54	0.005	
$TEMA\text{-}3_i = \beta_0 + \beta_1(NSB_i) + \beta_2(TEN\ Composite_i) + \beta_3(ENT_i) + \varepsilon_i$					
Constant	14.94	1.93	7.74	<0.001	
NSB	0.75	0.17	4.50	<0.001	0.64/ (0.03)
TEN Composite	0.04	0.02	1.91	0.03	
ENT	0.18	0.16	1.16	0.125	

The first model, which included both the NSB and the TEN Composite as predictors, explained 64% of the variance in TEMA-3 scores. This reflected a 3% change in R^2 values, and demonstrated that the combination of the NSB and the TEN Composite predicted performance on the TEMA-3 better than the NSB alone. In other words, in this model, the TEN Composite explained a significant amount of variation in TEMA-3 scores that was not already accounted for by the NSB ($p = 0.002$). The next model, which included the NSB and the ENT as predictors, predicted TEMA-3 performance above and beyond that of just the NSB, as well ($R^2 = 0.63$; R^2 change = 0.02). Similar to the first model, the ENT in this model explained a significant amount of variation in TEMA-3 scores that was not already explained by the NSB ($p = 0.005$). The final model, which included all three measures of number sense, also predicted TEMA-3 scores above that of just the NSB, although a closer examination of the results reveals some interesting findings. The R^2 and the change in R^2 values for this model were 0.64 and 0.03,

respectively. These values were no different than the values obtained from the first model, which contained only the NSB and TEN Composite. Consequently, the addition of the ENT to the NSB and TEN Composite did not explain the variance in TEMA-3 scores any better than the combination of those two measures alone ($p = 0.125$). The R^2 values for the third model were, however, slightly different than those obtained in the second model containing the NSB and ENT (R^2 values differed by 0.01). The addition of the TEN Composite to the NSB and ENT explained significantly more of the variance in TEMA-3 scores than just the NSB and ENT alone ($p = 0.03$).

While the above analyses answer the proposed research question, they do not reveal which early numeracy skills are associated with each assessment or which specific number sense components may be explaining the variance in TEMA-3 scores. Consequently, a content analysis of each number sense assessment was conducted in order to identify the overlapping and unique number sense components measured by each test (Table 11).

Table 11: Content analysis of each number sense assessment.

Number Sense Component	TEN	NSB	ENT
Oral counting	X	X	X
Quantity discrimination	X	X	X
Number identification	X	X	
Missing number	X		
One-to-one correspondence		X	X
Non-verbal computation		X	X
Verbal computation		X	X
Knowledge of number line		X	
Counting principles		X	
Skip counting			X
Counting on			X
Counting backwards			X
Resultative counting			X
Concepts of comparison			X
Classification			X
Seriation			X
Ordinality			X
Subitizing			X

Of the 18 different skills measured by the assessments, only two of the skills – oral counting and quantity discrimination – were found on all three measures. Number identification was shared by the TEN and NSB, while one-to-one correspondence, nonverbal computation, and verbal computation were skills shared by the NSB and ENT. In terms of skills that were unique to each test, missing number was found only on the TEN. Knowledge of the number line and counting principles were unique to the NSB. The ENT assessed the broadest range of skills and had nine unique components ranging from skip counting to seriation to subitizing.

The previous multiple regression analyses showed that combinations of the NSB with the other number sense measures explained more variance in TEMA-3 scores, above and beyond that of just the NSB. This additional explained variance could be due to the fact that the TEN and ENT each measure unique components that the NSB does not. For

example, adding the TEN to a model with the NSB resulted in a 3% change in R^2 values. It is possible that the missing number task, which is the one unique component of the TEN, was responsible for this additional explained variance. To investigate this further, an exploratory multiple regression analysis was conducted to determine which component(s) of the TEN were responsible for the change in R^2 values. In this analysis, the NSB was entered into the model first, followed by all four subtests of the TEN: Oral Counting (OC), Number Identification (NI), Quantity Discrimination (QD), and Missing Number (MN). The Holm-Bonferroni sequential procedure was then used to correct for multiple comparisons and control for family-wise error. Results indicated that in addition to the variance explained by NSB, OC and MN also explained a significant amount of the variance in TEMA-3 scores (p = 0.005 and 0.011, respectively), while QD and NI did not (p = 0.15 and 0.34, respectively; Table 12).

Table 12: Multiple regression analysis using NSB and TEN subtests as predictor variables.

	Unstandardized Coefficients					
Predictor	β	SE	t	p	$\alpha = 0.05/n$	R^2/ (R^2 Change)
$TEMA\text{-}3_i = \beta_0 + \beta_1(NSB_i) + \beta_2(OC_i) + \beta_3(NI_i) + \beta_3(QD_i) + \beta_4(MN_i) + \varepsilon_i$						
Constant	16.61	2.09	7.93	<0.001		
NSB	0.72	0.16	4.40	<0.001		
OC	0.12	0.04	2.67	0.005	0.0125	0.66/
MN	0.34	0.15	2.33	0.011	0.0167	(0.05)
QD	-0.11	0.10	-1.04	0.15	0.025	
NI	0.02	0.05	0.43	0.34	0.05	

Research Question Five

Does performance on a number sense measure, combined with fall teacher rating of number sense, predict mathematics achievement above and beyond that of performance on a number sense measure alone?

A multiple regression analysis was used to ascertain whether the combination of a number sense measure, plus fall teacher rating of number sense, would predict TEMA-3 performance above that of the number sense measure alone. It was hypothesized that the combination of the two variables would predict later mathematics achievement above that of performance on a number sense measure alone. Once again, because the NSB was considered to be the best number sense measure in a qualitative sense, one regression model was created in which the NSB was entered into the model first, followed by fall teacher rating of number sense. Results indicated that the addition of the fall teacher rating to the model resulted in a 1% change in R^2; however, fall teacher rating of number sense did not explain a significant amount of additional variance in TEMA-3 scores ($p = 0.14$; Table 13). The change in R^2 was most likely simply due to another predictor variable being added to the model.

Table 13: Multiple regression analysis combining NSB with fall teacher rating of number sense ($TEMA\text{-}3_i = \beta_0 + \beta_1(NSB_i) + \beta_2(Fall\ Rating_i) + \varepsilon_i$).

	Unstandardized Coefficients				
Predictor	β	SE	t	p	R^2/ (R^2 Change)
Constant	12.87	2.22	5.79	<0.001	0.62/ (0.01)
NSB	1.12	0.11	10.47	<0.001	
Fall T. Rating	0.47	0.32	1.49	0.14	

Research Question Six

Is there a relationship between a teacher's rating of a kindergartener's number sense in the spring and that same kindergartener's mathematics achievement in the spring?

A simple linear regression analysis was conducted to determine if there was a significant relationship between spring teacher rating of number sense and spring performance on the TEMA-3. It was hypothesized that there would be a positive

relationship between spring teacher rating of number sense and performance on the TEMA-3. Analyses revealed a moderately strong, positive relationship between spring teacher rating of number sense and spring mathematics achievement ($R = 0.65$; Table 14). Additionally, results indicated the regression model was significant ($p < 0.001$), suggesting that concurrent validity exists and that spring teacher rating of number sense significantly predicts spring performance on the TEMA-3. The R^2 value also showed that spring teacher rating explained 42% of the variability in TEMA-3 scores.

Table 14: Simple linear regression analysis between TEMA-3 and spring teacher rating of number sense ($TEMA\text{-}3_i = \beta_0 + \beta_1(Spring\ rating_i) + \varepsilon_i$).

	Unstandardized Coefficients						95% Confidence Interval for β	
Predictor	β	SE	R	R^2	t	p	Lower	Upper
Spring rating	2.76	0.32	0.65	0.42	8.53	<0.001	2.12	3.40

CHAPTER 5

DISCUSSION

The overall purpose of this study was to investigate the predictive validity of three measures of number sense – the Test of Early Numeracy (TEN), Number Sense Brief Screener (NSB), and Early Numeracy Test (ENT) – and to establish which measure, if any, best predicts later mathematics achievement. This study was also designed to determine if a particular combination of number sense assessments predicts future mathematics achievement significantly better than any one assessment alone. The relationship between performance on each number sense measure and teacher rating of number sense was examined, as well. In addition, the final purpose of this study was to add to the literature base on the psychometric properties of the TEN, NSB, and ENT by replicating research that has previously been conducted on these measures.

Summary of Findings

This study provided solid evidence for the predictive validity of the TEN, NSB, and ENT. As hypothesized, performance on all three number sense measures in the fall of kindergarten significantly predicted performance on a measure of mathematics achievement, the Test of Early Mathematics Ability, Third Edition (TEMA-3), in the spring of kindergarten. In fact, analyses revealed strong positive relationships between each number sense measure and later performance on the TEMA-3, as standardized regression coefficients ranged from 0.71 to 0.78. These findings are consistent with those of previous studies and suggest that number sense assessments, administered at the very beginning of kindergarten, are indeed predictive of future success in mathematics. Prior research on the TEN has demonstrated that performance on the kindergarten measures is

predictive of later performance on a range of other assessments such as the Number Knowledge Test, Stanford 10 Achievement Test, and mathematics CBM probes (Baglici et al., 2010; Chard et al., 2005; Martinez et al., 2009). The same is true for the NSB, as previous research has shown that kindergarten performance on the NSB predicts third grade performance on a high-stakes state test and on standardized measures of mathematics achievement, such as the Math Achievement subtest of the Woodcock-Johnson III (Jordan et al, 2008; Jordan, Glutting, Ramineni, & Watkins, 2010). Scores on the NSB at the beginning of first grade were also predictive of scores on the mathematics subtests of the Woodcock-Johnson III at the end of first and third grades (Jordan, Glutting, & Ramineni, 2010). Less evidence exists for the predictive validity of the ENT, although kindergarten scores on the Finnish version of the assessment were predictive of overall first grade mathematics performance (Aunio & Niemivirta, 2010). This study was the first to provide evidence for the predictive validity of the English version of the ENT.

Although correlations between each number sense measure and the TEMA-3 differed slightly, there were no significant differences in the way each measure predicted later mathematics achievement. In other words, no number sense measure emerged as the best predictor of future success in mathematics – the TEN, NSB, and ENT all predicted performance on the TEMA-3 similarly. This finding was inconsistent with one of the hypotheses of this study, as it was believed that the ENT would emerge as the best predictor of mathematics achievement. Of the three number sense measures utilized in this study, the ENT was the most comprehensive in that it assessed the broadest range of number sense components. These results demonstrate, however, that a more comprehensive number sense assessment may not always result in a better screening

instrument. The TEN, for example, only assesses four components of number sense. Despite this brevity, results indicated that the TEN predicted later mathematics achievement just as well as the ENT, which assesses over ten different components. These findings suggest that it may not matter how many number sense components are assessed, but rather which number sense components are assessed.

While the dependent correlations between each number sense measure and the TEMA-3 did not differ significantly, it may be important to investigate the differences in the way the NSB and ENT each predicted TEMA-3 scores with a larger sample. In this study, the correlation between the NSB and the TEMA-3 was 0.78, and the correlation between the ENT and the TEMA-3 was 0.71. This difference in predictive validity was not significant, but did result in an observed p-value of 0.08. Consequently, this same comparison should be tested with a larger sample size, as more power could result in significant findings. Of course, even if statistical significance between the NSB and ENT were detected with a larger sample size, this would not necessarily demonstrate clinical significance. Practitioners may, for example, be less concerned with small statistical differences in predictive validity and more concerned with the qualitative features of the assessments, such as ease of administration and scoring.

In the case of this study, the qualitative features of each assessment were one of the only factors that set the assessments apart. While statistical analyses did not reveal any one measure as being the best predictor of later mathematics achievement, an informal review of each assessment resulted in the NSB emerging as the "best" measure in a qualitative sense. Anecdotal evidence from data collectors who administered all three measures (n = 6) suggested that of the three number sense assessments utilized in

this study, the NSB was considered the simplest to administer and easiest to score. In addition, data collectors reported that the NSB appeared to be the most engaging to students. Data collectors also noted that the TEN was more difficult to administer with integrity due to its timed nature, while the ENT was described as somewhat lengthy and tedious for young children.

Just as previous research on the TEN, NSB, and ENT had not compared the predictive validity of these three measures directly, no studies had examined the predictive validity of different combinations of these measures, either. Findings from this study indicate that certain combinations of number sense measures predict mathematics achievement above and beyond that of just one measure alone. Performance on the NSB and TEN, for example, predicted TEMA-3 performance better than performance on the NSB alone. The combination of the NSB and TEN together explained 64% of the variance in TEMA-3 scores, while the NSB alone explained 61% of the variance in TEMA-3 scores. Similar results were found with the combination of the NSB and ENT, as these two measures together explained variation in TEMA-3 scores better than the NSB alone (R^2 change = 0.02). A subsequent content analysis, which identified the overlapping and unique skills assessed by each number sense measure, suggested that this additional explained variance was most likely due to the fact that each measure assesses at least one unique early numeracy skill that the others do not. The TEN, for example, is the only measure that assesses students' ability to identify the missing number in a string of three digits (e.g., 7 __ 9). The ENT measures fourteen different number sense components, nine of which are unique to the ENT and include skills such as subitizing, seriation, and classification.

While the content analysis suggested that Missing Number, the unique component of the TEN, was solely responsible for the additional variance explained by the combination of the TEN and NSB, a multiple regression analysis further investigating this notion revealed some interesting findings. When the NSB was placed into a regression model first, followed by all four subtests of the TEN, results showed that in addition to the variance explained by the NSB, both Missing Number and Oral Counting also explained a significant amount of the variance in TEMA-3 scores, while Number Identification and Quantity Discrimination did not. Consistent with the conclusions made from the content analysis, the multiple regression analysis demonstrated that in this model, Missing Number explained a significant amount of the variance in TEMA-3 scores. Interestingly, Oral Counting also significantly contributed to the change in R^2, even though counting aloud was a skill already measured by the NSB. This particular finding is likely due to the fact that the skill of oral counting is assessed in greater depth and is weighted much more heavily on the TEN than on the NSB. On the TEN, students are given one minute to count as high as they can, with 100 as the limit. Their score is the highest number they reach at one minute, minus any errors (e.g., if a student counts to 47 in one minute but makes one error, their oral counting score would be 46). On the NSB, however, there is only one item that addresses oral counting. On this item, students are asked to count to ten. If they do this successfully, they receive one point; if they are unable to count to ten successfully, they receive zero points. Thus, the ability to count aloud represents approximately one quarter of the items on the TEN, compared to one thirty-third of the items on the NSB. In addition, the TEN allows for a more precise measurement of the skill, while the NSB only assigns an all-or-nothing score.

Unfortunately, this type of skill analysis was not possible to conduct with the ENT or NSB, as these assessments do not consist of separate subtests, but rather only contain one or two items that measure each number sense component. Regardless, these results again support the notion that it may not matter how many number sense components are assessed, but rather which number sense components are assessed and at what depth.

Finally, it should also be noted that the combination of all three number sense measures did not explain variation in TEMA-3 scores any better than the combination of the NSB and TEN. The combination of all three measures did, however, explain variation in TEMA-3 scores slightly better than the combination of the NSB and ENT. When placed in a model with the NSB and TEN, the ENT did not explain any significant variance in TEMA-3 scores. The addition of the TEN to a model already containing the NSB and ENT, however, did explain significantly more of the variance in TEMA-3 scores than just the NSB and ENT alone.

These results were somewhat consistent with this study's hypotheses. Since a combination of two or three measures would assess an increasingly broader range of number sense components, it was hypothesized that each combination of number sense measures created would predict mathematics achievement above that of just one measure alone. Findings supported this hypothesis in that combining the NSB with one other measure of number sense resulted in a model that predicted later mathematics achievement better than the NSB alone. Findings were contrary to this hypothesis in that adding a third measure to the model did not always further explain the variability in TEMA-3 performance. The ENT, for example, does not appear to measure any unique components that contribute to later mathematics achievement beyond that already

assessed by the combination of the NSB and TEN. The addition of the TEN to the NSB and ENT, however, does create a model that explains significantly more of the variance in TEMA-3 scores.

Once again, the issue of clinical significance must be kept in mind when considering these findings. Although the combinations of the NSB and TEN and NSB and ENT explained performance on the TEMA-3 better than the NSB alone, these combinations of two number sense measures only explained two to three percent more of the variance in TEMA-3 scores (e.g., NSB alone explained 61% of the variance in TEMA-3 scores, while the combination of the NSB and TEN explained 64% of the variance). While this change is statistically noteworthy, one has to wonder about its clinical significance or practical importance. Educators are not likely to administer two entire number sense assessments simply because administering two assessments predicts later mathematics achievement slightly better than one assessment. Instead, practitioners are more likely to administer one assessment that is reliable, valid, simple, and brief.

In addition to investigating the predictive validity of different measures of number sense, this study also examined the concurrent and predictive validity of teacher rating of number sense. Consistent with hypotheses, a significant positive relationship was found between teacher rating of number sense in the spring and TEMA-3 scores, thereby providing evidence for the concurrent validity of teacher rating of student number sense. Teachers appear to have a fairly strong understanding of their students' number sense skills in the spring of kindergarten, as their rating of student number sense explained 42% of the variation in TEMA-3 scores. Teacher rating of student number sense in the fall also significantly predicted spring performance on the TEMA-3; however, the fall rating

only explained a relatively small percentage (18%) of the variance in TEMA-3 scores. This was again in line with this study's hypotheses. Further examination of these results revealed that fall teacher rating of number sense did not predict later mathematics achievement as well as the number sense measures. In fact, all three of the number sense assessments predicted TEMA-3 performance significantly better than fall teacher rating of number sense. These results are not entirely surprising, as teachers completed the fall number sense ratings during the first two weeks of school and were not very familiar with their students or their ability levels.

Furthermore, analysis of a model containing both the NSB and fall teacher rating of number sense showed that fall teacher rating did not explain a significant amount of variance in TEMA-3 scores. This was contrary to hypotheses, as it was believed that teacher rating would add a unique component to the model not captured by a number sense assessment. Although having teachers rate their students' number sense skills at the very beginning of kindergarten would likely be better than the absence of any screening exercise, findings indicate that using a specific number sense assessment such as the TEN, NSB, or ENT would more accurately identify students who are at risk for struggling in mathematics. In addition, because combining number sense measures did not yield clinically meaningful changes in predictive validity, the use of one assessment to screen for mathematics difficulties appears to be sufficient. Although all three number sense measures predicted mathematics achievement similarly in a statistical sense, anecdotal evidence garnered from data collectors suggests that the NSB may be the "best" measure in a qualitative sense. This measure was reportedly the easiest to administer and score, and also appeared to be the most engaging to students.

Limitations

A number of important limitations must be considered when interpreting the results of this study. The design of this particular study involved the administration of several number sense assessments over a short period of time (two weeks) at the very beginning of the school year. Multiple administrations of similar assessments may have threatened the internal validity of these results, as some students may have, for example, performed better on the ENT simply because they were exposed to similar questions on the TEN and NSB in the week prior, thus potentially skewing the predictive validity of the ENT. The administration of the three assessments in the fall were purposely not counterbalanced in order to ensure that the amount of time between each assessment was the same for each participant; however, counterbalancing may have mitigated the potential effects of repeated testing. Although students took each number sense test on a separate day, repeated testing could have also contributed to fatigue for some students. Fatigue on the latter tests (the NSB and ENT) may have influenced student performance on these assessments and consequently affected the findings regarding the tests' predictive utility.

The fact that some students received additional instruction in mathematics beyond what was offered in the general education classroom is another notable limitation to the current study. Over the course of the school year, nine students received intervention support in mathematics and four received specialized mathematics instruction via special education services. While these students only represent approximately 10% of the overall sample, it is possible that this additional mathematics instruction may have

contributed to higher TEMA-3 scores for these students, thus affecting the predictive validity of the three number sense measures.

An additional threat to the results of this study may be attributed to the attrition of ten students from the original sample of 112 kindergarteners. Seven of these ten students moved out of the district mid-year, two students were non-English speakers, and one student chose not to participate. The exclusion of these students from the final sample may have influenced the overall results of the study. For example, if these students had remained in the sample, their inclusion may have provided analyses with more power to detect a smaller difference in the way each measure predicted later mathematics achievement. It is possible, however, that the results might have remained the same even with the inclusion of these ten students.

The way in which participants were recruited may have threatened the validity of this study, as well. In order for students to participate, parents were required to complete and return a consent form describing the study. In a way, this created a self-selected sample of kindergarteners. It is possible that there was a difference between those parents who allowed their children to participate and those parents who chose not to have their children participate.

One final limitation, and perhaps the most significant threat to the external validity of this study, was the resulting composition of this study's sample. Participants in this study were from two mid-sized, suburban elementary schools whose populations were not exceptionally ethnically or socioeconomically diverse. In both schools, over 85% of the student population was Caucasian and approximately one third of the student population was eligible for free or reduced-price lunch. In addition, although a

significant portion of these schools' student population is comprised of English language learners, few of these students were given consent to participate in the study. As a result, the findings of this study may not be generalizable to schools that have more diverse student populations.

Implications for Practice

While the need for reliable and valid measures of number sense and early numeracy skills is clear, research on the psychometric properties and predictive utility of these measures is arguably still in the preliminary stages. The results of this study not only add to the growing literature base on the assessment of number sense and early numeracy skills, but they also provide important implications for practice. Consistent with prior research (Aunio & Niemivirta, 2010; Baglici et al., 2010; Chard et al., 2005; Jordan et al., 2008; Jordan, Glutting, Ramineni, and Watkins, 2010; Martinez et al., 2009), performance on the TEN, NSB, and ENT at the beginning of kindergarten predicted end-of-kindergarten performance on a standardized measure of mathematics achievement. None of these measures, however, emerged as the "best" predictor of later mathematics achievement; the TEN, NSB, and ENT all predicted future success in mathematics similarly. This information is valuable for practitioners for a number of reasons. First, these results provide educators with assurance that the TEN, NSB, and ENT are all valid screening instruments that assess skills predictive of later success in mathematics. Consequently, these three measures can all be used to identify students who may be at risk for experiencing later difficulties in mathematics. The use of these measures for early identification purposes would then hopefully result in early intervention services for those students identified as at risk. In addition, knowing that no

single number sense measure emerged as the best predictor of later mathematics achievement allows for practitioners to use the assessment that best fits their needs and the needs of their schools. Educators who work with large student populations and are pressed for time might choose to use the TEN due to its brief administration time, while those practitioners interested in assessing a wider range of skills might use the NSB or ENT.

While statistical analyses revealed that certain combinations of number sense measures predicted mathematics achievement above that of just one measure, the difference in the way two measures predicted achievement versus one was clinically insignificant. The small amount of information that a second measure might provide does not appear to outweigh the time and effort it would take to administer multiple assessments, let alone justify the extra time the student would be out of the classroom for testing purposes. As a result, it is advisable that educators only give students one number sense measure during fall screenings, as the results of one assessment are likely to provide as much information as administering two or three measures.

Although findings clearly support the use of either the TEN, NSB, or ENT for the purposes of screening students for mathematics difficulties, a closer examination of the results from this study provides practitioners with a window into the number sense skills most important for students at kindergarten entry. As discussed, the number and depth of skills assessed on the TEN, NSB, and ENT varies greatly. The ENT, for example, broadly assesses fourteen different number sense components, while the TEN measures four components in relatively greater depth. Because all three number sense measures predict later mathematics achievement in the same way, it seems that kindergarteners'

performance on the four skills measured by the TEN – oral counting, number identification, quantity discrimination, and missing number – gives practitioners just as much information as their performance on the fourteen skills measured by the ENT. In other words, students who have a solid grasp of a few very important skills at kindergarten entry (e.g., those skills that appear on the TEN) will most likely perform just as well on an end-of-kindergarten mathematics assessment as students who have a solid understanding of an incredibly broad range of early numeracy skills (e.g., subitizing, classification, seriation, ordinality). As a result, practitioners should focus on helping their pre-kindergarten students build a solid, in-depth understanding of a few basic early numeracy skills, rather than a partial understanding of several early numeracy skills. This recommendation is consistent with National Mathematics Advisory Panel's (2008) observation that mathematics is a hierarchical subject area whereby more complex skills are built on simpler, foundational skills. In preschool and kindergarten, it appears most important to foster the development and mastery of foundational skills such as counting and number identification rather than attempt to teach a broad range of skills that may not contribute as significantly to later mathematics achievement.

Finally, in terms of teacher perception of student number sense, findings from this study suggest that when making educational decisions for a student, teachers should not rely solely on their perception of that student's number sense skills. Although fall teacher rating of student number sense was predictive of mathematics achievement in the spring, analyses showed that the TEN, NSB, and ENT all predicted mathematics achievement significantly better than fall teacher rating. In addition, the combination of a number sense measure (i.e., the NSB) with fall teacher rating did not predict later

mathematics achievement much better than the number sense measure alone. Ultimately, these results demonstrate the value of administering assessments such as the TEN, NSB, and ENT that are specifically designed to measure number sense skills and that accurately predict later mathematics achievement. Depending on teacher perception of student number sense alone is not likely to yield an accurate picture of which students may experience mathematics difficulties in the future.

Directions for Future Research

The research base on number sense within the field of education – and on early mathematics skills as a whole – has grown substantially over the last several years (Gersten et al., 2011). Despite this recent growth, additional research is needed in a number of key areas in order to continue expanding on what is currently known about early mathematics assessment and achievement. Although the current study revealed a number of unique and interesting results, replicating the study with a larger sample that is more representative of the overall population would likely increase the generalizability of the results. In addition, while studies investigating the predictive validity of a measure over the course of a year are useful, more longitudinal research is needed. It would, for example, be interesting to determine if the TEN, NSB, and ENT predict mathematics achievement at the end of elementary school and even at the end of middle or high school. Jordan and her colleagues (2008) did demonstrate that kindergarten performance on the NSB predicted third grade performance on a standardized mathematics achievement test, but no studies have looked at the predictive validity of the measure beyond third grade. Similarly, longitudinal research on the predictive qualities of the TEN and ENT has not been conducted. Replicating the current study using other, less

widely used measures of number sense may also be valuable in identifying an assessment that is the "best" predictor of later mathematics achievement. Several additional measures of number sense and early numeracy currently exist, and inclusion of these measures into a similar study may result in one or two measures emerging as the better predictors of future success in mathematics.

Specific research on the predictive validity of different number sense components is needed, as well. In other words, future studies should focus on which components of number sense have the strongest relationship with later mathematics achievement. Research in this area could reveal, for example, that proficiency in one-to-one correspondence is extremely predictive of later mathematics achievement, while simple computation is not. Knowledge of which components are most closely related to later mathematics achievement would better inform the development and refinement of new and current number sense assessments.

While additional research on the predictive validity of different number sense components and assessments is necessary and useful, further investigation into the classification accuracy of these measures is also needed. As Gersten et al. (2012) explain, classification accuracy is the "degree to which the screener provides correct classifications of children who require additional assistance" (p. 437). Instruments with a high level of classification accuracy are both sensitive and specific. Measures that are sensitive consistently identify students who actually need extra academic intervention, and measures that are specific do not misclassify students as needing intervention when in fact they do not (Gersten et al., 2012). Number sense measures that have demonstrated predictive validity can only provide educators with a potential level of risk for any given

student; predictive validity cannot, however, tell educators which at-risk students will definitely develop difficulties in mathematics without intervention and which students will go on to succeed without any intervention (Gersten et al., 2012). Thus, the need for determining the classification accuracy, sensitivity, and specificity of each number sense assessment is clear, yet the methods for examining classification accuracy are relatively new to educational research. Recently, researchers in the field of early mathematics assessment have begun using receiver operating characteristic (ROC) curve analyses to examine classification accuracy and to identify cut scores for assessments that strike an ideal balance between sensitivity and specificity (Gersten et al., 2012; Jordan, Glutting, Ramineni, & Watkins, 2010). Of the three number sense measures utilized in this study, researchers have investigated the classification accuracy of the NSB and TEN (Jordan, Glutting, Ramineni, & Watkins, 2010; National Center on Response to Intervention, 2014). An initial study examining the way the NSB predicted scores on a high-stakes state test showed that the screener had a fairly high degree of classification accuracy (AUC, which represents the area under the curve, was 0.80 when the NSB was administered in the fall of kindergarten) (Jordan, Glutting, Ramineni, & Watkins, 2010). The National Center on Response to Intervention (2014) has also reported preliminary information on the classification accuracy of the TEN when predicting later performance on curriculum-based measures in mathematics. When the four subtests of the TEN were administered in in the fall of kindergarten, classification accuracy was generally good, as AUC values ranged from 0.85 to 0.87. The Center's *Screening Tools Chart*, however, denotes that there is "unconvincing evidence" regarding the classification accuracy of the TEN (National Center on Response to Intervention, 2014). Consequently, additional

research on the classification accuracy of the TEN and other measures needs to be conducted. Once more is understood about the classification accuracy of various number sense measures, practitioners will have even more useful information for determining which measure to use in the screening of early mathematics skills.

Of course, any future research on the assessment of number sense should be accompanied by research on interventions designed to improve the number sense and early numeracy skills of young children identified as at risk. A small collection of these types of interventions currently exist, including Number Worlds (Griffin, 2004), numerical board games (Ramani & Siegler, 2008; Siegler, 2009), and explicit instruction in concepts such as number recognition, sequencing, and problem solving (Jordan, Glutting, Dyson, Hassinger-Das, & Irwin, 2012). Each of these interventions has shown promise in fostering the number sense skills of young children, but further examination of their effectiveness is needed. The development of a wider variety of evidence-based number sense interventions is imperative, as well. As researchers begin to identify and construct number sense assessments with high levels of predictive validity and classification accuracy, evidence-based interventions targeting both broad and specific components of number sense need to be readily available to practitioners so that they may help support the needs of students at risk for future difficulties in mathematics.

APPENDIX A
PARENT/GUARDIAN CONSENT FORM

Dear parent or guardian,

My name is Bethany Politylo and I am currently a doctoral student in the School Psychology program at the University of Massachusetts Amherst. My research interests lie in mathematics education, and as part of my dissertation, I am planning to administer four different mathematics assessments to the kindergarteners in your school district. I am writing to request permission for your child to participate in my research study.

About the study
Much of the research in early mathematics education has highlighted the importance of a child's "number sense" or early numeracy skills for future success in mathematics. Several measures that assess a child's number sense currently exist, but none have been very well researched. The purpose of my study is to analyze three different number sense assessments and determine which, if any, best predicts future success in mathematics. My hope is to find the number sense assessment that can best identify students who are struggling or succeeding in mathematics very early on; this way, teachers can provide the necessary and appropriate support to ensure that these students are successful in mathematics later in life.

Assessments
The three number sense assessments I am using in my study are individually administered and relatively brief. The first assessment takes about 10 minutes to complete and involves counting, identifying numbers, simple calculations, and comparing quantities. The second assessment takes about 20-25 minutes to complete and entails classifying objects, ordering from big to small, counting, and working with shapes. The final number sense assessment involves similar tasks and takes only four minutes to complete. This final assessment is one that the staff in your school district already administers to all kindergarteners.

The three assessments described above will be administered in late September. In May, I will give one final assessment of mathematics achievement. This assessment is also individually administered, and takes about 30 minutes to complete. Skills assessed include counting, identifying and manipulating numbers, and simple calculations. All of the above assessments are essentially extensions of the activities and assessments that generally already occur in kindergarten.

Confidentiality
All assessment materials and scores gathered as part of this study will be kept strictly confidential. Your school district will have access to student scores on the four-minute number sense assessment described above, as this is the assessment that the district normally administers to all kindergarteners. All other assessment materials and scores will only be reviewed by myself and members of my dissertation committee. In addition, because I am interested in overall performance on these assessments by a large group of

kindergarteners, your child's individual scores will not be analyzed. All data from this study will be compiled into a large data set; the data set will not contain student names, and as a result, there is no risk of your child's scores being identified. Any publications that result from this research will report the overall performance of all participants. No student's individual scores will be singled out and examined. Results of the study will be made available to your school district.

Participation
Participation in this study is completely voluntary, and your child will not be penalized in any way if he or she does not participate. In addition, you have the right to withdraw your permission for your child to participate at any time.

Please fill out the form below and have your child return it to his or her classroom teacher by Friday, September 14.

If you have questions or concerns regarding this study, please feel free to contact me at bpolityl@educ.umass.edu. You may also contact Dr. Amanda Marcotte, my advisor and dissertation chairperson, at 413-545-7055 or amarcotte@educ.umass.edu.

Sincerely,

Bethany Politylo
Doctoral Candidate, School Psychology
University of Massachusetts Amherst
bpolityl@educ.umass.edu

My child, _____, **has / does not have** (circle one) permission to participate in the mathematics research study described above.

_____ _____
Parent/guardian signature Date

APPENDIX B
TEACHER RATING SCALE

Dear _____,

The National Mathematics Advisory Panel (2008) defines number sense as "… an ability to immediately identify the numerical value associated with small quantities (e.g., 3 pennies), a facility with basic counting skills, and a proficiency in approximating the magnitudes of small numbers of objects and simple numerical operations."

The Panel also adds that "an intuitive sense of the magnitudes of small whole numbers is evident even among most 5-year-olds who can, for example, accurately judge which of two single digits is larger, estimate the number of dots on a page, and determine the approximate location of single digit numerals on a number line that provides only the numerical endpoints. These competencies comprise the core number sense that children often acquire informally prior to starting school.

A more advanced type of number sense that children must acquire through formal instruction requires a principled understanding of place value, of how whole numbers can be composed and decomposed, and of the meaning of the basic arithmetic operations of addition, subtraction, multiplication, and division. It also requires understanding the commutative, associative, and distributive properties and knowing how to apply these principles to solve problems. This more highly developed form of number sense should extend to numbers written in fraction, decimal, percent, and exponential forms."

On a scale of 1 to 10, with 1 representing a poorly developed number sense and 10 representing a well-developed number sense, please rate the current number sense of _____ to the best of your ability by circling the appropriate number below.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Poorly developed number sense									*Well-developed number sense*

Thank you!

APPENDIX C
ADMINISTRATION INTEGRITY CHECKLISTS

Oral Counting Accuracy in Implementation Rating Scale (AIRS)

Oral Counting Accuracy in Implementation Rating Scale (AIRS)

Examiner: _____ Date: Observation 1 _____
Observer: _____ Observation 2 _____
Observation 3 _____

X = completed accurately O = incorrect

Step	Observation 1	Observation 2	Observation 3
Seated appropriate distance from child			
Places examiner copy out of view of child			
Says standardized directions			
Turns tape recorder on (optional)			
Says "Start"			
Starts stopwatch at correct time (after student says first number)			
Marks errors on examiner copy			
Times accurately for 1 minute			
Says "Stop"			
Stops stopwatch			
Marks last number stated with a bracket			
Turns off tape recorder (optional)			
Determines # of Correct Oral Counts			
Records score			

Number Identification Accuracy in Implementation Rating Scale (AIRS)

Number Identification Accuracy in Implementation Rating Scale (AIRS)

Examiner: _____ Date: Observation 1 _____
Observer: _____ Observation 2 _____
Observation 3 _____

X = completed accurately O = incorrect

Step	Observation 1	Observation 2	Observation 3
Seated appropriate distance from child			
Places practice item in front of child			
Places student copy in front of child			
Places examiner copy out of view of child			
Says standardized directions			
Turns tape recorder on (optional)			
Says "Start"			
Starts stopwatch at correct time (after student says first number)			
Marks errors on examiner copy			
Times accurately for 1 minute			
Says "Stop"			
Stops stopwatch			
Marks last number stated with a bracket			
Turns off tape recorder (optional)			
Determines # of Correct Number Identifications			
Records score			

Quantity Discrimination Accuracy in Implementation Rating Scale (AIRS)

Quantity Discrimination Accuracy in Implementation Rating Scale (AIRS)

Examiner: _____ Date: Observation 1 _____
Observer: _____ Observation 2 _____
Observation 3 _____

X = completed accurately O = incorrect

Step	Observation 1	Observation 2	Observation 3
Seated appropriate distance from child			
Places practice item in front of child			
Places student copy in front of child			
Places examiner copy out of view of child			
Says standardized directions			
Turns tape recorder on (optional)			
Says "Start"			
Starts stopwatch at correct time (after student says first number)			
Marks errors on examiner copy			
Times accurately for 1 minute			
Says "Stop"			
Stops stopwatch			
Marks the last item completed with a bracket			
Turns off tape recorder (optional)			
Determines number of Correct Quantity Discriminations			
Records score			

Missing Number Accuracy in Implementation Rating Scale (AIRS)

Missing Number Accuracy in Implementation Rating Scale (AIRS)

Examiner: _____ Date: Observation 1 _____
Observer: _____ Observation 2 _____
Observation 3 _____

X = completed accurately O = incorrect

Step	Observation 1	Observation 2	Observation 3
Seated appropriate distance from child			
Places practice item in front of child			
Places student copy in front of child			
Places examiner copy out of view of child			
Says standardized directions			
Turns tape recorder on (optional)			
Says "Start"			
Starts stopwatch at correct time (after student says first number)			
Marks the last item completed with a bracket			
Times accurately for 1 minute			
Says "Stop"			
Stops stopwatch			
Marks last letter or words read with a bracket			
Turns off tape recorder (optional)			
Determines number of Correct Missing Numbers			
Records score			

Your name: _____ Name of primary data collector: _____

Student name: _____ Classroom: _____ Date + time: _____

Number Sense Brief Screener (NSB)

Administration integrity checklist

Please use the below checklist to assess the administration integrity of the NSB. Do not discuss or share this checklist with primary data collector.

Timeline	**Step**	**X = completed accurately; O = completed incorrectly or did not complete**
Before testing	Seated across from child	
	Places scoring sheet out of view of child	
Items 1-2	Gives standardized instructions for items 1-2	
	Shows child picture of 5 stars	
Item 3	Gives standardized instructions for item 3	
	Stops child after he/she reaches 20 (if applicable)	
Items 4-7	Gives standardized instructions for items 4-7	
	Shows child picture of 5 dots	
	Counts left to right with puppet	
	Counts right to left with puppet	
	Counts yellow dots then blue with puppet	
	Counts first dot twice with puppet	
Items 8-11	Gives standardized instructions for items 8-11	
	Shows child numbers on separate pieces of paper	
Items 12-18	Gives standardized instructions for items 12-18	
Items 19-22	Gives standardized instructions for items 19-22	
	Gives child 2 pieces of card stock and 10 chips	
	Models each item by adding/removing chips one at a time	
Items 23-33	Gives standardized instructions for items 23-33	
After testing	Tallies number of correct responses	
	Records score	

Your name: _____ Name of primary data collector: _____

Student name: _____ Classroom: _____ Date + time: _____

Early Numeracy Test (ENT)

Administration integrity checklist

Please use the below checklist to assess the administration integrity of the ENT. Do not discuss or share this checklist with primary data collector.

Timeline	**Step**	**X = completed accurately; O = completed incorrectly or did not complete**
Before testing	Seated across from child	
	Places scoring sheet out of view of child	
	Introduces test using standardized introduction	
Items 1-5	Gives standardized instructions for items 1-5	
	Shows child appropriate probes in binder for items 1-5	
Items 6-10	Gives standardized instructions for items 6-10	
	Shows child appropriate probes in binder for items 6-10	
Items 11-15	Gives standardized instructions for items 11-15	
	Shows child appropriate probes in binder for items 11-15	
	Gives child 10 blocks for item 11	
	Gives child 15 blocks for item 12	
	Gives child paper and pencil for item 13	
	Gives child paper pencil for item 14	
Items 16-20	Gives standardized instructions for items 16-20	
	Shows child appropriate probes in binder for items 16-20	
	Gives child paper and pencil for item 19	
Items 21-25	Gives standardized instructions for items 21-25	
	Shows child appropriate probes in binder for items 22 + 24	
Items 26-30	Gives standardized instructions for items 26-30	
	Lays down 16 blocks in four rows of four blocks for item 26	
	Lays down 9 blocks in a circle for item 27	
	Lays down 20 blocks in a heap for item 28	

	Shows child appropriate probe in binder for item 29	
	Lays down 17 blocks in a row for item 30	
Items 31-35	Gives standardized instructions for items 31-35	
	Gives child 15 blocks for item 31	
	Lays down 20 blocks in a row for item 32	
	Lays down 15 blocks in three rows of five for item 33	
	Lays down 19 blocks in a heap for item 34	
	Does not allow child to point out or touch blocks for items 32-34	
	Covers five blocks with hands then adds seven for item 35	
Items 36-40	Gives standardized instructions for items 36-40	
	Shows child appropriate probes in binder for items 36-40	
After testing	Tallies number of correct responses	
	Records score	

Your name: _____ Name of primary data collector: _____

Student name: _____ Classroom: _____ Date + time: _____

TEMA-3

Administration integrity checklist

Please use the below checklist to assess the administration integrity of the TEMA-3. Do not record administration integrity for items that were not administered (i.e., student reaches basal/ceiling and thus those items need not be administered). Do not discuss or share this checklist with primary data collector.

Timeline	**Step**	**X = completed accurately; O = completed incorrectly or did not complete**
Before testing	Seated across from or next to child	
	Places scoring sheet out of view of child	
	Asks child how old s/he is and begins at appropriate item	
Items A1-A6 (if applicable)	Gives standardized instructions for items A1-A6	
	Shows child appropriate probes in Picture Book for items A1, A4, + A6	
	Uses 12 tokens and 3 cards for item A5	
Items A7-A14 (if applicable)	Gives standardized instructions for items A7-A14	
	Shows child appropriate probes in Picture Book for items A7 and A14	
	Uses 12 tokens and 3 cards for item A8	
	Uses 5 tokens for item A9	
	Uses 10 tokens for item A10	
	Uses 10 tokens for item A12	
Items A15-A21 (if applicable)	Gives standardized instructions for items A15-A21	
	Gives child worksheet and pencil for item A15	
	Uses 10 tokens for item A16	
	Uses 10 tokens for item A17	
	Shows child appropriate probes in Picture Book for item A18	

Items A22-A31 (if applicable)	Gives standardized instructions for items A22-A31	
	Shows child appropriate probes in Picture Book for items A23, A27, and A29	
	Uses 12 tokens for item A25	
	Uses 10 tokens for item A25, but **does NOT** let child use them	
	Uses 25 tokens for item A28	
	Gives child worksheet and pencil for item A30	
Items A32-A42 (if applicable)	Gives standardized instructions for items A32-A42	
	Shows child appropriate probes in Picture Book for items A35, A37, A38, and A41	
	Gives child worksheet and pencil for item A34	
	Covers up problems quickly (after a couple seconds) for item A41	
Items A43-A56 (if applicable)	Gives standardized instructions for items A43-A56	
	Shows child appropriate probes in Picture Book for items A43, A44, A46, A47, A50, A51, A53-A56	
	Covers up problems quickly (after a couple seconds) for item A43	
	Gives child worksheet and pencil for item A45	
	Covers up problems quickly (after a couple seconds) for item A46	
	Gives child worksheet and pencil for item A49	
	Covers up problems quickly (after a couple seconds) for item A50	
	Covers up problems quickly (after a couple seconds) for item A51	
	Covers up problems quickly (after a couple seconds) for item A54	
	Covers up problems quickly (after a couple seconds) for item A56	
Items A57-A72 (if applicable)	Gives standardized instructions for items A57-A72	
	Shows child appropriate probes in Picture Book for items A58-A61, A66, and A70	
	Covers up problems quickly (after a couple seconds) for item A61	
	Gives child worksheet and pencil for item A62 and A63	
	Gives child worksheet and pencil for item A63	
	Gives child worksheet and pencil for item A69	
	Covers up problems quickly (after a couple seconds) for item A70	
	Gives child worksheet and pencil for item A71	

Basal/ceiling rules	Discontinues testing after ceiling is reached (five consecutive scores of zero)	
	If necessary, after reaching discontinue criteria, tests backwards from starting point until basal is met (five items correct in a row) or until item A1 is reached	

APPENDIX D
CALCULATING A TEN COMPOSITE SCORE

Several methods for calculating a composite score for the Test of Early Numeracy (TEN) were considered and tested. These methods are as follows:

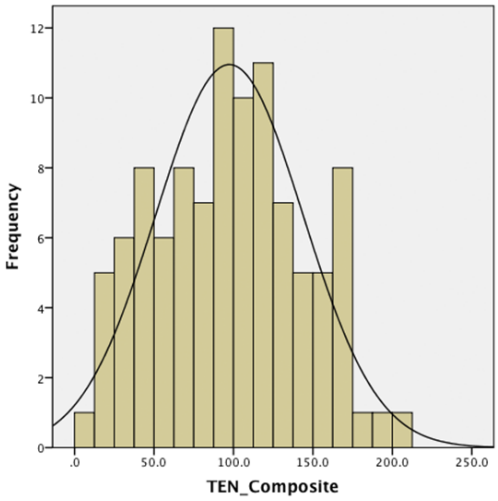
1. Sum the scores obtained on each subtest.
2. Convert each subtest score to a proportion of items correct. Sum the proportions.
3. Convert the score on each subtest to a *z*-score. Sum the *z*-scores.
4. Conduct a Principal Components Analysis (PCA). Use the resulting component score as the TEN composite score.

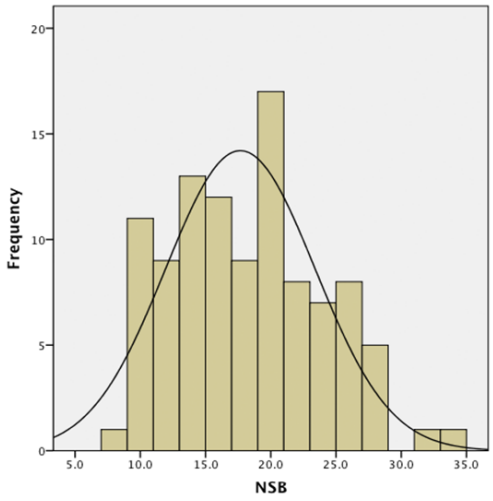
The below correlation matrix shows the relationship between each of these new composite scores as well as their relationship with the criterion measure, the TEMA-3.

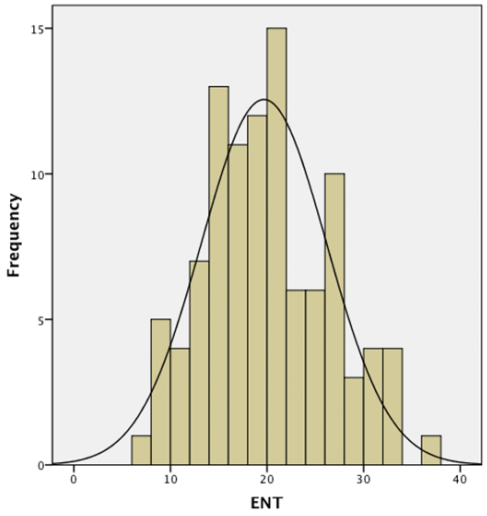
	TEN Sum	TEN Proportion	TEN *z*-score sum	TEN PCA	TEMA-3
TEN Sum	1.00	0.99	0.99	0.99	0.73
TEN Proportion		1.00	1.00	1.00	0.73
TEN *z*-score sum			1.00	1.00	0.74
TEN PCA				1.00	0.74
TEMA-3					1.00

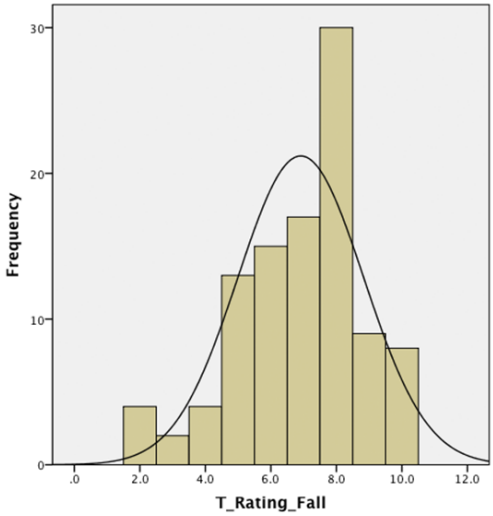
Given the strong relationship between each of the methods, the simplest method of summing each subtest score was used to create the TEN composite score in this study.

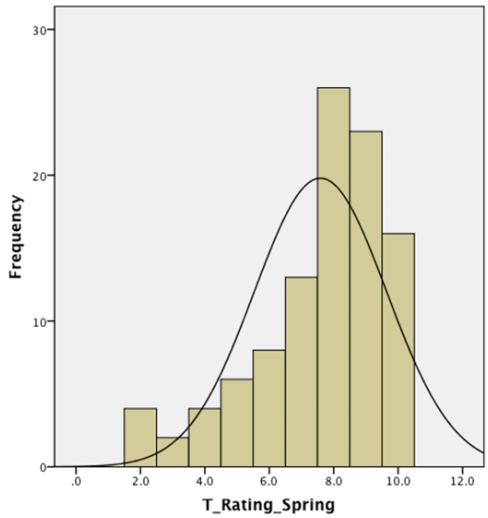
APPENDIX E
HISTOGRAMS OF INDEPENDENT AND DEPENDENT VARIABLES

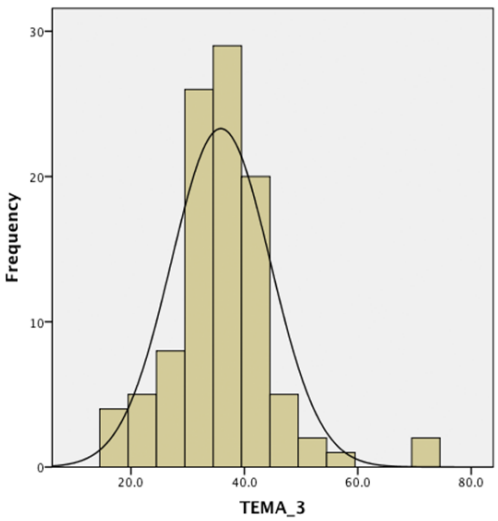




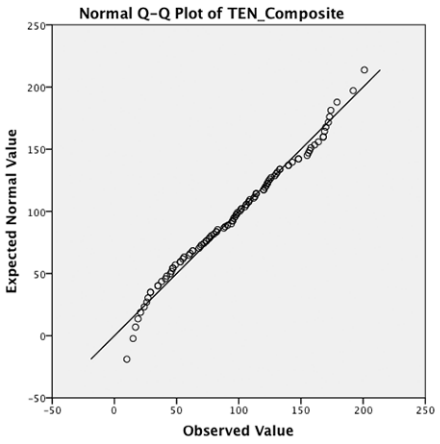


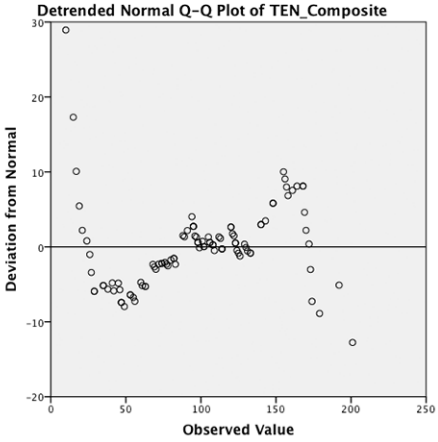


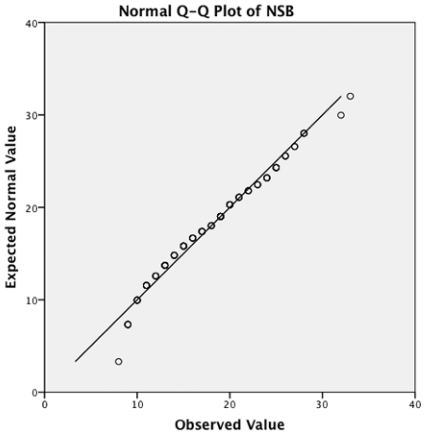


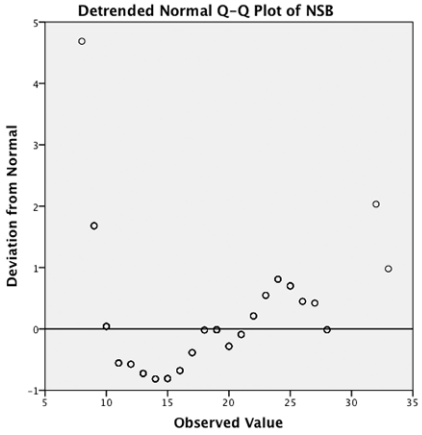


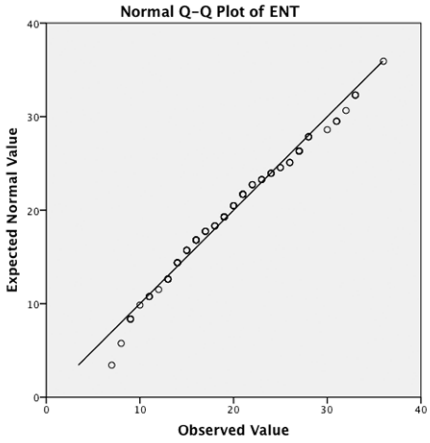
APPENDIX F
Q-Q PLOTS OF INDEPENDENT AND DEPENDENT VARIABLES

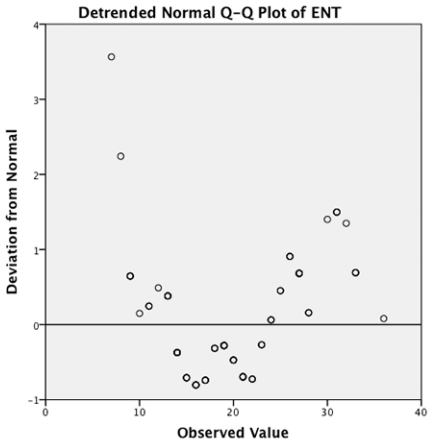


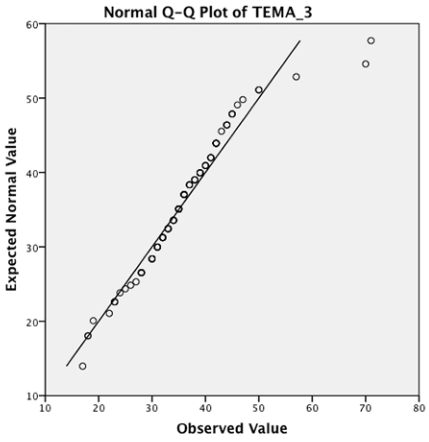


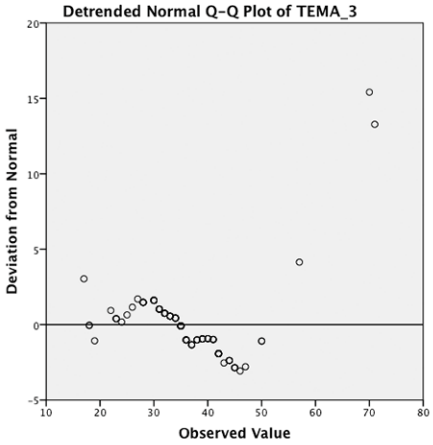


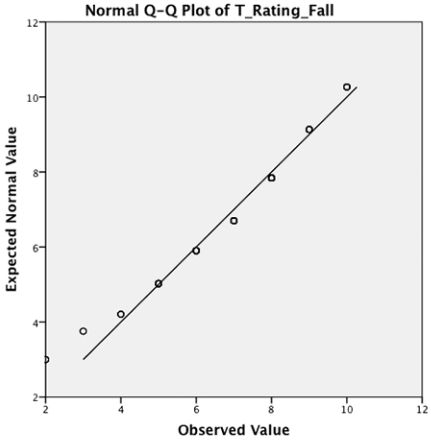


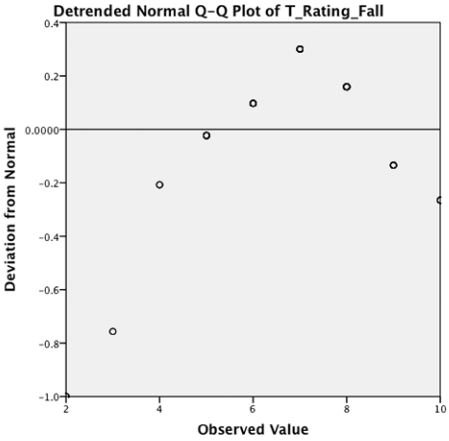


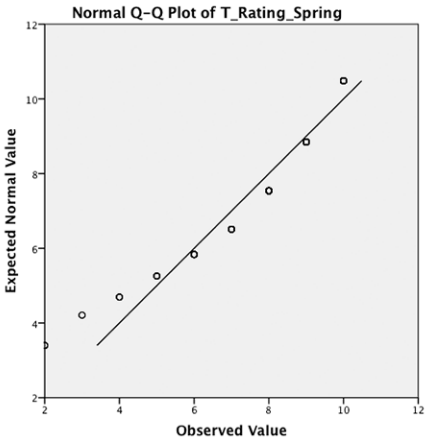


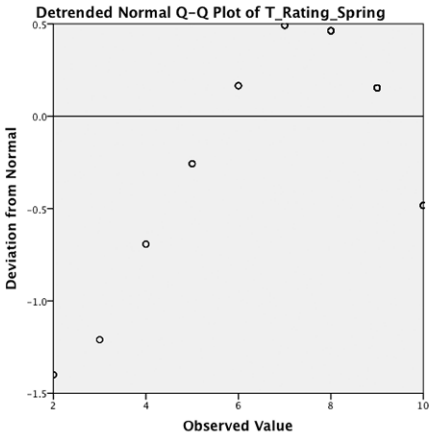

Normal Q-Q Plot of TEMA_3
Expected Normal Value
Observed Value


Detrended Normal Q-Q Plot of TEMA_3
Deviation from Normal
Observed Value

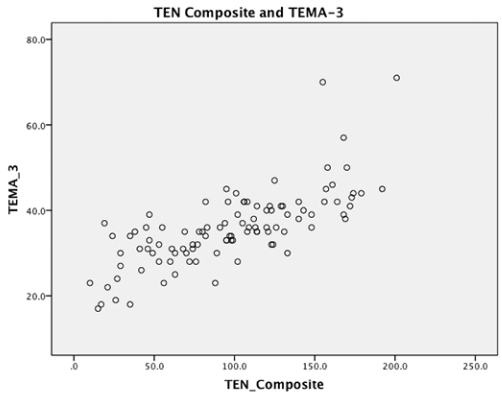

Normal Q-Q Plot of T_Rating_Fall
Expected Normal Value
Observed Value

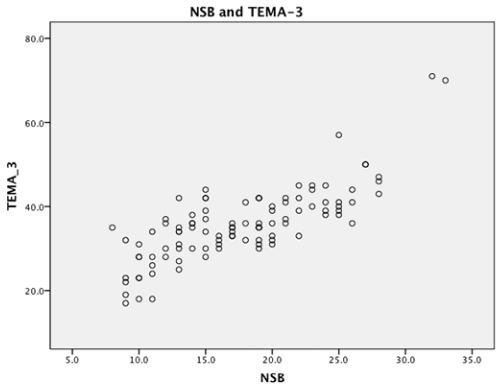

Detrended Normal Q-Q Plot of T_Rating_Fall
Deviation from Normal
Observed Value

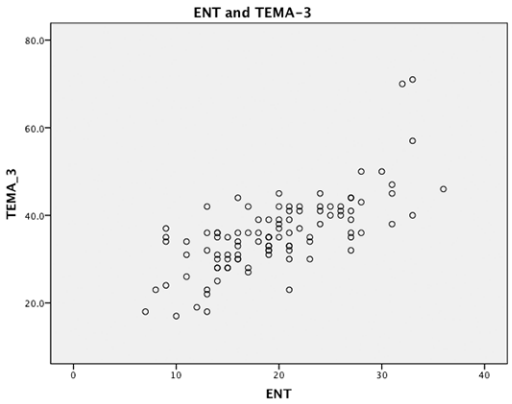

Normal Q-Q Plot of T_Rating_Spring
Expected Normal Value
Observed Value

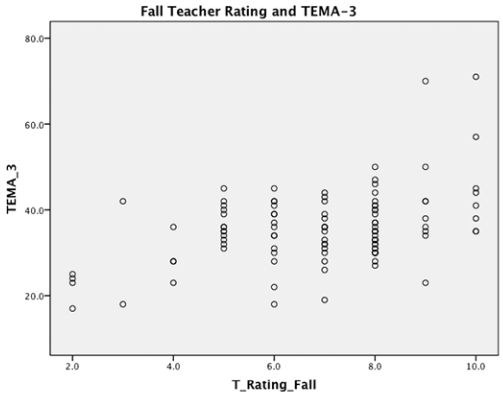

Detrended Normal Q-Q Plot of T_Rating_Spring
Deviation from Normal
Observed Value

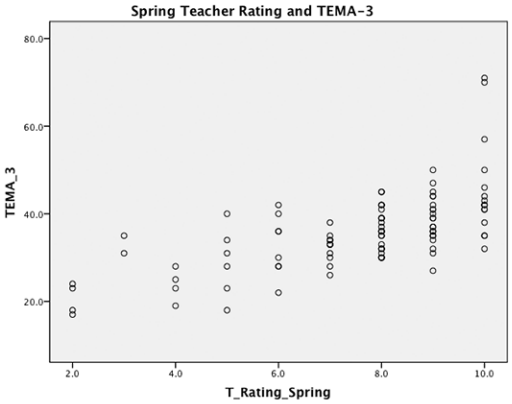
APPENDIX G
SCATTERPLOTS OF INDEPENDENT VARIABLES WITH TEMA-3



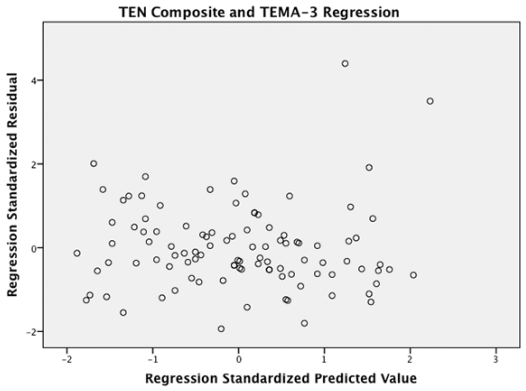


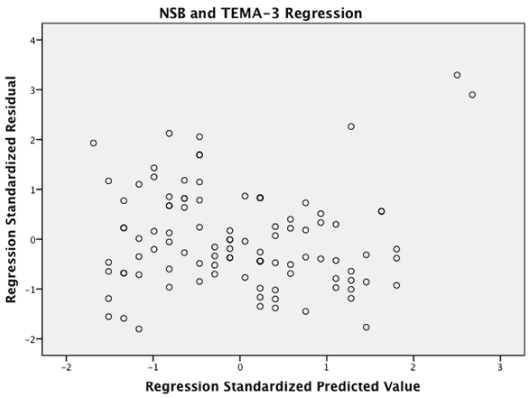


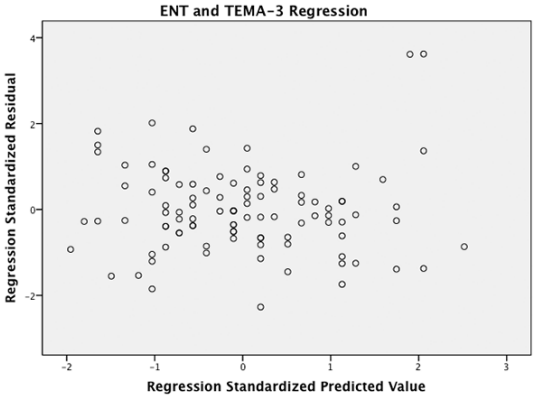


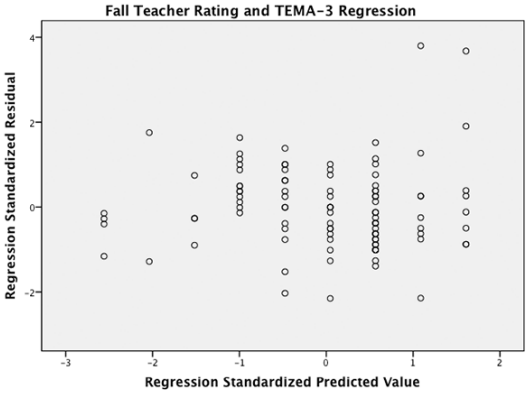


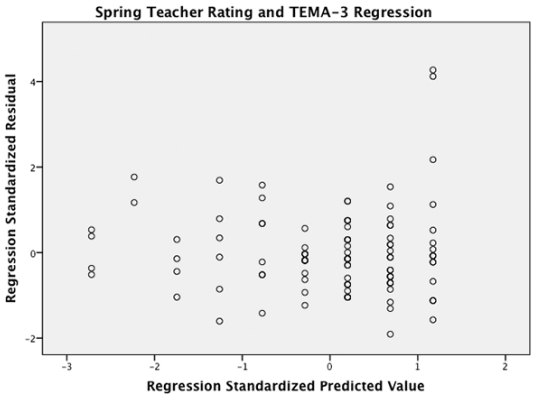
APPENDIX H
SCATTERPLOTS OF RESIDUALS VERSUS PREDICTED VALUES











REFERENCES

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. Eugene, OR: Center for Educational Assessment Accountability.

Aunio, P., Hautamäki, J., Heiskari, P., & Van Luit, J. E. H. (2006). The Early Numeracy Test in Finnish: Children's norms. *Scandinavian Journal of Psychology*, *47*, 369-378.

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*, 427-435.

Aunio, P., Niemivirta, M., Hautamäki, J., Van Luit, J. E. H., Shi, J., & Zhang, M. (2006). Young children's number sense in China and Finland. *Scandinavian Journal of Educational Research*, *50*(5), 483-502.

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the research on the Tests of Early Numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention*, *35*(2), 89-102.

Baker, S., Gersten, R., Katz, R., Chard, D., & Clarke, B. (2002). *Preventing mathematics difficulties in young children: Focus on effective screening of early number sense delays* (Tech. Rep. No. 0305). Eugene, OR: Pacific Institutes for Research.

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

Cantlon, J. F., & Brannon, E. M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science*, *17*(5), 401-406.

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., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement. *Remedial and Special Education*, *29*(1), 46-57.

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. L. M., Tindal, G., Kame'enui, E. J., & Baker, S. K. (2011). Classification accuracy of easyCBM first-grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention*, *36*(4), 243-255.

Clarke, B., & Shinn, M. R. (2004a). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*(2), 234-248.

Clarke, B., & Shinn, M. R. (2004b). *Test of Early Numeracy (TEN): Administration and scoring of AIMSweb early numeracy measures for use with AIMSweb*. Eden Prairie, MN: Edformation, Inc.

Clements, D. H. (1984). Training effects on the development and generalization of Piagetian local operations and knowledge of number. *Journal of Educational Psychology*, *76*(5), 766-776.

Dantzig, T. (1946). *Number: The language of science* (3rd ed.). New York, NY: The Macmillan Company.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2009). G*Power (Version 3.1.3) [Computer software]. Retrieved from http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Fernald, D. (1984). *The Hans legacy: A story of science*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gersten, R., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education*, *33*(1), 18-28.

Gersten, R., Clarke, B., Haymond, K., & Jordan, N. C. (2011). *Screening for mathematics difficulties in K-3 students* (2nd ed.). Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Council for Exceptional Children*, *78*(4), 423-445.

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

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability, Third Edition*. Austin, TX: Pro-Ed, Inc.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*, 496-499.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education*, *22*(3), 170-218.

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

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense:" The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457-1465.

Howden, H. (1989). Teaching number sense. *The Arithmetic Teacher*, *36*(6), 6-11.

Howell, S., & Kemp, C. (2005). Defining early number sense: A participatory Australian study. *Educational Psychology*, 2*5*(5), 555-571.

Howell, S., & Kemp, C. (2006). An international perspective of early number sense: Identifying components predictive of difficulties in early mathematics achievement. *Australian Journal of Learning Disabilities*, *11*(4), 197-207.

Howell, S., & Kemp, C. (2009). A participatory approach to the identification of measures of number sense in children prior to school entry. *International Journal of Early Years Education*, *17*(1), 47-65.

Ifrah, G. (1985). *From one to zero: A universal history of numbers*. New York, NY: Viking Penguin, Inc.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(25), 10382-10385.

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology*, *104*(3), 647-660.

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowder (Ed.), *Mathematical difficulties: Psychology and intervention* (pp. 45-58). San Diego, CA: Academic Press.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, *20*, 82-88.

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, *39*(2), 181-195.

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*(1), 36-46.

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in Kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*(1), 153-175.

Kelly, D., Xie, H., Nord, C.W., Jenkins, F., Chan, J.Y., and Kastberg, D. (2013). *Performance of U.S. 15-year-old students in mathematics, science, and reading literacy in an international context: First look at PISA 2012* (NCES 2014-024). Retrieved from http://nces.ed.gov/pubs2014/2014024rev.pdf

Kilian, A., Yaman, S., Von Fersen, L., & Güntürkün, O. (2003). A bottlenose dolphin discriminates visual stimuli differing in numerosity. *Learning & Behavior*, *31*(2), 133-142.

Koehler, O. (1951). The ability of birds to count. *The Bulletin of Animal Behaviour*, *9*, 41-45.

Lago, R. M., & DiPerna, J. C. (2010). Number sense in kindergarten: A factor-analytic study of the construct. *School Psychology Review*, *39*(2), 164-180.

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice*, *24*(1), 12-20.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292-1300.

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science*, *14*(5), 396-401.

Malofeeva, E., Day, J., Saco, X., Young, L., & Ciancio, D. (2004). Construction and evaluation of a number sense test with Head Start children. *Journal of Educational Psychology*, *96*(4), 648-659.

Martinez, R. S., Missall, K. N., Graney, S. B., Aricak, O. T., & Clarke, B. (2009). Technical adequacy of early numeracy curriculum-based measurement in kindergarten. *Assessment for Effective Intervention*, *34*(2), 116-125.

Massachusetts Department of Elementary and Secondary Education. (2013). *School/District profiles*. Retrieved from http://profiles.doe.mass.edu/

McComb, K., Packer, C., & Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, *Panthera leo*. *Animal Behavior*, *47*, 379-387.

Mcintosh, A., Reys, B. J., & Reys, R. E. (1992). A proposed framework for examining basic number sense. *For the Learning of Mathematics*, *12*, 2-8.

Mechner, F. (1958). Probability relations within response sequences under ratio reinforcement. *Journal of the Experimental Analysis of Behavior*, *1*, 109-121.

Methe, S. A., Hojnoski, R., Clarke, B., Owens. B. B., Lilley, P. K., Politylo, B. C., … Marcotte, A. M. (2011). Innovations and future directions for early numeracy curriculum-based measurement: Commentary on the special series. *Assessment for Effective Intervention*, *36*(4), 200-209.

National Center for Education Statistics. (2013). *The nation's report card: A first look: 2013 mathematics and reading* (NCES 2014-451). Retrieved from http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014451.pdf

National Center on Response to Intervention. (2014). *Screening tools chart*. Retrieved from http://www.rti4success.org/resources/tools-charts/screening-tools-chart

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

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Retrieved from http://www.nichd.nih.gov/publications/pubs/nrp/documents/report.pdf

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Retrieved from http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: Mathematics Learning Study Committee.

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. In R. Case & Y. Okamoto (Eds.), *The role of central conceptual structure in the development of children's thought: Monographs of the Society for Research in Child Development* (Vol. 1-2, pp. 27-58). Malden, MA: Blackwell Publishers.

Pepperberg, I. M. (2006). Grey parrot (*Psittacus erithacus*) numerical abilities: Addition and further experiments on a zero-like concept. *Journal of Comparative Psychology*, *120*(1), 1-11.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., … Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*, 33-41.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*, 499-503.

Politylo, B. C., White, K. M., & Marcotte, A. M. (2011, February). *An investigation of the construct of number sense*. Poster session presented at the meeting of the National Association of School Psychologists, San Francisco, CA.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). *Highlights from TIMSS 2011: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2013-009 Revised). Retrieved from http://nces.ed.gov/pubs2013/2013009_1.pdf

R Development Core Team. (2012). R: A language and environment for statistical computing (Version 2.15.0) [Computer software]. Retrieved from http://www.R-project.org

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, *79*(2), 375-394.

Revelle, W. (2013). Psych: Procedures for personality and psychological research (Version 1.3.2) [Computer software]. Retrieved from http://CRAN.R-project.org/package=psych

Reys, R., Reys, B., Mcintosh, A., Emanuelsson, G., Johansson, B., & Yang, D. C. (1999). Assessing number sense of students in Australia, Sweden, Taiwan, and the United States. *School Science and Mathematics*, *99*(2), 61-70.

Rilling, M. (1993). Invisible counting animals: A history of contributions from comparative psychology, ethology, and learning theory. In S. T. Boysen & E. J. Capaldi (Eds.), *The development of numerical competence: Animal and human models* (pp. 3-37). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Siegler, R. S. (2009). Improving the numerical understanding of children from low-income families. *Child Development Perspectives*, *3*(2), 118-134.

Sowder, J. T. (1989). Introduction. In J. T. Sowder & B. P. Schappelle (Eds.), *Establishing foundations for research on number sense and related topics: Report of a conference* (pp. 1-5). San Diego, CA: Center for Research in Mathematics Education, San Diego State University.

Sowder, J. T. (1992). Making sense of numbers in school mathematics. In G. Leinhardt, R. Putman, & R. A. Hattrup (Eds.), *Analysis of arithmetic for mathematics teaching* (pp. 1-51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Starkey, P., & Cooper, R. G. (1980). Perception of numbers by human infants. *Science*, *210*, 1033-1035.

Starkey, P., Spelke, E. S., & Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. *Science*, *222*, 179-181.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18116-18120.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Van de Rijt, B. A. M., Godfrey, R., Aubrey, C., Van Luit, J. E. H., Ghesquière, P., Torbeyns, J.,…Tzouriadou, M. (2003). The development of early numeracy in Europe. *Journal of Early Childhood Research*, *1*(2), 155-180.

Van de Rijt, B. A. M., Van Luit, J. E. H., & Pennings, A. H. (1999). The construction of the Utrecht Early Mathematical Competence Scales. *Educational and Psychological Measurement*, *59*, 289-309.

Van Luit, J. E. H., & Van de Rijt, B. A. M. (2005). *Early Numeracy Test* (3rd ed.). Doetinchem, The Netherlands: Graviant Publishing Company.

Woodruff, G., & Premack, D. (1981). Primitive mathematical concepts in the chimpanzee: Proportionality and numerosity. *Nature*, *293*, 568-570.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749-750.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*, B1-B11.

Yang, D. C., Hsu, C. J., & Huang, M. C. (2004). A study of teaching and learning number sense for sixth grade students in Taiwan. *International Journal of Science and Mathematics Education, 2*, 407-430.