Investigating the Effects of the MathemAntics Number Line Activity on Children's Number Sense

Samantha Creighan

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
Under the Executive Committee
Of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Investigating the Effects of the MathemAntics Number Line Activity on Children's Number Sense

Samantha Creighan

Number sense, which can broadly thought of as the ability to quickly understand, approximate, and manipulate numerical quantities, can be a difficult construct for researchers to operationally define for empirical study. Regardless, many researchers agree it plays an important role in the development of the symbolic number system, which requires children to master many tasks such as counting, indentifying numerals, comparing magnitudes, transforming numbers and performing operations, estimating, and detecting number patterns, skills which are predictive of later math achievement. The number line is a powerful model of symbolic number consistent with researchers' hypotheses concerning the mental representation of number. The MathemAntics Number Line Activity (MANL) transforms the number line into a virtual manipulative, encourages estimation, provides multiple attempts, feedback, and scaffolding, and introduces a novel features where the user can define his own level of risk on the number line. The aim of the present study was to examine how these key features of MANL are best implemented to promote number sense in low-income second-graders. Sixty-six students from three schools were randomly assigned to one of three conditions; MANL User-Defined Range (UDR), and MANL Fixed Range (FR), and a Reading comparison condition and underwent a pretest session, four computer sessions, and a posttest session. During the computer sessions, researchers coded a child's observed strategy in placing targets on the number line. The results showed that children with higher number sense ability at pretest performed better on a posttest number line estimation measure when they were in the UDR condition than in the FR condition. Conversely, children with low number sense ability at pretest performed better on the number line estimation posttest measure when they were in the FR condition than UDR. Although in general, all children improved over time, children with low number sense ability at pretest were more likely to use the UDR tool ineffectively, thus negatively impacting performance. When children were not coded as responding quickly, target number significantly impacted performance in the computer sessions. Finally, children in the UDR condition utilized better expressed strategies on the number line estimation posttest than children in the Reading comparison group. These findings indicate that prior number sense ability plays a role in how children engage with MANL, which in turn affects the learning benefits the child receives. Implications for researchers, software designers, and math educators, as well as limitations are discussed.

TABLE OF CONTENTS

Chapter 3: Results

Chapter 4: Discussion

List of Figures

List of Tables

ACKNOWLEDGEMENTS

There are many people who provided help and support during the completion of this research project. First and foremost I would like to express my gratitude to Herb Ginsburg for providing me with the opportunity of being involved in the MathemAntics project and his outstanding advisement and support throughout my time at Teachers College. I would also like to thank my dissertation committee, Ryan Baker, Gary Natriello, Janet Metcalfe, and Tom Hatch for their thoughtful feedback and discussions. I am grateful to the MathemAntics research team including Ama Awotwi, Kara Carpenter, Azedah Jamalian, Rachael Labrecque, Dana Pagar, and Esther Yoon. Our weekly team meetings and dissertation club sessions were invaluable to the completion of this project. Additionally, thanks to Koleen McCrink and Dor Abrahamson who provided feedback on the design of various MathemAntics activities, including the one used in this study. I would also like to thank the many patient and flexible data collectors for bearing with me during the course of the study, and to the school administrators, teachers, students, and parents who participated in my study.

I could not have completed my Ph.D. without my parents, who have instilled in me a lifelong love of learning and who along with the rest of my family have been a source of unwavering support and unconditional love. Finally, I want to thank my husband Jon, who has played so many roles throughout the completion of this dissertation including test subject, project manager, equipment handler, personal chef, therapist, and best friend.

DEDICATION

For Beans.

Chapter 1

INTRODUCTION

The Nature of Number Sense

Number Sense allows us to make quick and intuitive judgments about the relations among numbers. For example, it helps us use basic ideas to avoid drudgery (such as the commutative principle, which teaches us that 5 + 4 is the same as 4 + 5, thus eliminating the need to calculate or remember the second sum if you already know the first), quickly break numbers into convenient parts, and have a general "feel" for numbers, such as knowing 185 + 141 cannot be smaller than 200. It's easy to recognize number sense at work when you see it. Imagine enjoying dinner at a restaurant with a group of friends. When the bill comes, the group must calculate the tip and divide the bill evenly among the group. While some are still reaching for their smartphones or scribbling the formal algorithm on paper, someone has already figured it out mentally, perhaps by using a rounding strategy.

Although researchers agree that number sense exists and is important, the difficulty rises in operationally defining it in a way that allows it to be measured and studied empirically (Berch, 2005; Siegler, 2004). Brain imaging, comparative, and psychophysical data have lead some numerical cognition researchers to adopt a strict definition of number sense as purely non-symbolic quantity representation, and thus argue it utilizes the approximate number system (ANS).

Mathematical thinking develops well before a child even learns to speak. Preverbal infants and nonhuman animals can represent and identify exact quantities of up to 3 or 4 quickly and accurately, a phenomenon known as subitizing (Feigenson, Dehaene & Spelke, 2004). Infants also possess the ability to recognize simple number transformations (Wynn, 1995). This exact small number system is paired in both humans and non-human animals with an approximate number system (ANS) that can represent larger quantities, allowing for approximate magnitude comparisons and calculations (Xu & Spelke, 2000; McCrink & Wynn, 2004; Gallistel, 1990). These two preverbal systems develop naturally without explicit training, and involve individual differences (Dehaene, 1997; Halberda & Feigenson, 2008).

The Development of Understanding Symbolic Number

Children begin to learn basic ideas about number in infancy and continue to develop this understanding throughout the elementary years (Mix, Huttenlocher, & Levine, 2002; Starkey, Spelke et al. 1990; Xu & Spelke, 2000). The concept of number is at the crux of many mathematical abilities for children. It is no surprise that the Common Core Standards (National Governors Association Center for Best Practices and Council of Chief State School Officers 2010), most curricula, and most educators (Cross, Woods, & Schweingruber, 2009) emphasize the importance of number in their mathematics standards and learning materials. Fortunately, the field of cognitive psychology has shed light on the complex cognitive processes involved in the development of the understanding of number, and stresses that a sophisticated concept of number requires the development of many systems, as I now describe.

Preverbal non-symbolic number competencies such as being able to identify small numbers and compare larger quantities, when the difference is larger than a specific ratio, provide the foundation for learning the symbolic number system, the ability to identify and represent exact quantities. In fact, many (Butterworth & Reeve, 2008; LeCorre, Van de Walle, Brannon, & Carey 2006; Dehaene, 1997; Gallistel & Gelman, 2005) argue that it is this conception of exact quantities that makes us mathematically superior to our ancestor and non-human mammal counterparts. Moyer and Landauer (1967; 1973) famously posited that humans mentally represent exact quantities as magnitudes on a mental number line. In a series of studies they asked adults to quickly judge the numerical order of two Arabic numerals, and found evidence for what they termed the distance and size effects. That is, adults take longer to judge the order of two numbers when these values are closer together (the distance effect), and they exhibit more difficulty with judgments for larger numbers than smaller (see also Dehaene, 1997; Gallistel et al.; 2005). This, coupled with recent findings that direct training on the ANS can improve performance on later symbolic number tasks (Hyde, Khanum, & Spelke, 2014) suggest that the ANS plays a role in mastering the symbolic number system.

Although the basis for many important mathematical ideas (such as more and less, one-to-one correspondence, and rudimentary approximation) occurs without the knowledge of the formal symbolic number system, humans must use symbolic number words to keep track of exact quantity. Knowledge of number words and the ability to use them for counting sets of objects is arguably an essential skill for daily navigation of adult human life in developed countries.

Counting and enumeration are early mathematical activities that expose children to this symbolic number system. In a seminal publication and book, Gelman and Gallistel (1986, 2005) outlined five basic counting principles: (1) the stable order principle, emphasizing the consistency of the number sequence, (2) the one-to-one correspondence principle, in which every object in the counting set is tagged with only one counting symbol, (3) the cardinal principle, which states that the number of the last object counted represents the total set value, (4) the abstraction principle, or the concept that one can count various types of entities as part of the same set, and (5) the order-irrelevance principle, which states that objects can be counted in any order. While some have argued these principles do not need to be mastered in order to enumerate

(LeCorre et al., 2006), repeated experience with counting is valuable and adds to children's development of understanding of precise number and cardinality, which is cognitively challenging for young children.

As children become better at enumerating larger sets, they must also begin to understand the base ten structure of the symbolic number system. Rather than counting by ones, a child is taught to group by tens, and must learn that numbers are comprised of tens and ones (Nuerk & Williams, 2005). Understanding the base-ten system and place value is an important mathematical acquisition for children, and this understanding lays the foundation for more-advanced topics such as multi-digit operations (Baroody, 1990). Despite its importance, many children struggle to master place value in the early elementary years. In particular, English-speaking children struggle with the numbers 11-19 due to the numbers' linguistic opacity (Li, 2002; Sun & Zhang, 2010). For example, the number word "twelve" gives no indication that it is comprised of one ten and two ones. Conversely, the Chinese word for twelve literally translates to "one ten and two ones", leading to better understanding of the teens.

Not only must children use preverbal number abilities to aid in mastering the symbolic number system, and integrate single-digit and multi-digit numbers to a unified base-ten conceptual system, but they must also utilize domain-general cognitive skills. As with any other fundamental and broad set of concepts, mathematical abilities do not develop independently from other cognitive functions. Cognitive researchers have discovered positive relationships between math abilities and literacy and language (Duncan, et al., 2007; Blair & Razza, 2007), executive functioning (Bull & Scerif, 2001), and spatial abilities (Tversky, 2011; Cook, Yip, Goldin-Meadow, 2012). The findings have also been supported by neuroscience research showing that although very basic number competencies occur in a separate area of the brain, advanced mathematical problem solving (e.g., word problems, abstract calculations) includes areas of the brain typically associated with higher order cognitive skills (Dehaene, Piazza, Pinel, & Cohen, 2003).

As discussed, understanding of number requires the integration of the preverbal and complex symbolic number systems, and integration with domain-general cognitive skills. Children must master many skills such as counting, indentifying numerals, comparing magnitudes, transforming number and performing operations, estimating, and detecting number patterns, (Ginsburg, 1989; Jordan, Kaplan, Oláh, & Locuniak, 2006.) These basic number competencies are predictive of later mathematics achievement, further highlighting the importance of fundamental mathematical concepts for young children (Dyson & Jordan, 2011; Libertus, Fiegenson, & Halberda, 2011). It is easy to conceptualize how a strong number sense, or better *number sense access*, the ability to quickly and often automatically mentally map symbolic number representations to a non-symbolic quantity, can aid in mathematical problem solving in many of these competencies in children as young as kindergarten (Dehaene & Akhavein, 1995; Gilmore et al., 2007; Wilson, Dehaene, Pinel, Revkin, Cohen, L, & Cohen, D., 2006). Stronger number sense access facilitates quick judgments about numbers and their relationships. For example, a child with strong number sense or number sense access can quickly determine that 9 is larger than 4 by five units, and can quickly estimate the sum of 34 and 57 to be somewhere close to 90. A child lacking this number sense and number sense access may rely more heavily on the formal algorithms and procedures, while lacking a strong conceptual understanding of why the algorithms work. Thus, many researchers have adopted a much broader definition of number sense as the "ability to quickly understand, approximate, and manipulate quantities" (Wilson, Dehaene, Dubois, & Fayol, 2006) without specifying the specific cognitive systems involved. Researchers have found that number sense defined by this broad definition is predictive of later math achievement (Jordan et al., 2006).

The lack of a consistent operational definition of number sense has resulted in some researchers developing and studying separate, but often related constructs (Dehaene, 1997; Jordan et al., 2006; Siegler & Booth, 2004). One common thread throughout many of the existing definitions is that number sense (or number sense access) involves knowledge of numbers, their magnitudes, and how their magnitude relates to other numbers. Thus, much of Siegler's work has focused on the study of "numerical magnitude representation," or how we mentally represent numerical magnitudes.

He (as well as others) argues that we represent numerical magnitudes on a series of mental number lines that are compressed logarithmically (e.g., more space on our mental number line is devoted to smaller numbers and larger numbers are compressed closer together). With further exposure to larger numbers and enumeration, the mental number line- that is, the mapping of quantities to a mental spatial continuum- becomes more accurate (Siegler & Booth, 2004). The number line estimation task, in which a child is given a blank number line with the endpoints marked and asked to place various target numbers on the line, is used to quantify participants' mental representation of number and has inspired a large body of research (Barth & Paladino, 2011; Booth & Siegler, 2006; Laski & Siegler, 2007; Moeller, Pixner, Kaufmann & Nuerk, 2008). By plotting the actual number versus the participant's response and determining the function of best fit, researchers can determine whether a participant is utilizing a logarithmic or linear representation of number for the given range (see Figure 1).

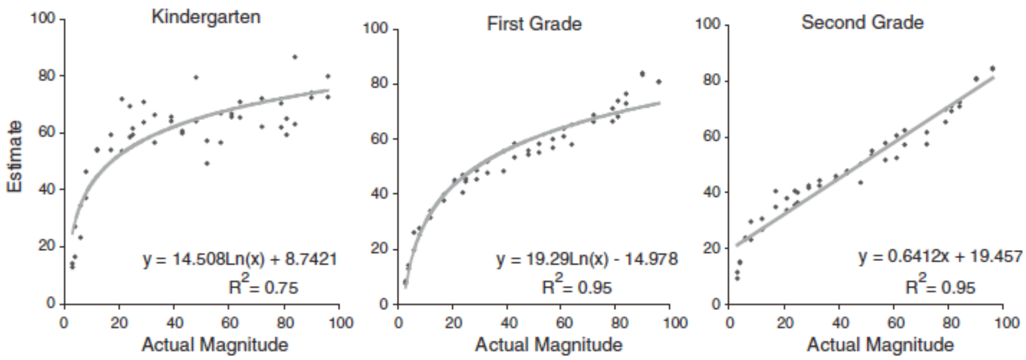

Figure 1: Demonstrates the log to linear shift in the number line estimation task (from Siegler & Booth, 2004)

Instruction and exposure to larger numbers, and feedback on performance at critical number values, help children shift from a logarithmic representation to a more accurate linear mental number line. Siegler & Booth (2004) found that instruction of this type may have long-term benefits; the likelihood of using a linear representation of number is highly predictive of later mathematical achievement (Jordan et al., 2006). Although there is some recent work suggesting that the number line task may not reflect a fundamental conceptual shift (Barth & Paladino, 2011, Opfer, Siegler, & Young, 2011), its usefulness as a practical tool for assessment and instruction, as well as the link between number line representations and later ability, are well documented.

Petitto (1990) examined the strategies children use in the number line estimation task. She found that typically, children begin from zero and count with an inaccurate unit (e.g., using a very tiny unit to count from zero to find 9 on a number line marked from 0-10), to counting forward with an accurate unit, to counting forward or backward depending on the target number (e.g., backward for targets closer to the right endpoint as the nine example), to abandoning counting strategies for partitioning strategies, often splitting the number line at the midpoint to find the target. In addition to accuracy, children's strategy use can reveal useful insights into their number sense abilities.

As well as assisting with the log to linear shift in numerical magnitude representation, the number line estimation task also involves estimation, which has also been linked to later math achievement (Jordan, et al., 2006). The log to linear shift has also been demonstrated when participants must estimate length, area, and volume (Booth & Siegler, 2006). Given the every-day frequency of approximation (e.g, estimating how much of my salary I spend on groceries or quickly guessing how long it will take me to drive from New York City to Pittsburgh), it is no surprise that Nancy Jordan and colleagues have clearly defined estimation as a critical component of number sense (Jordan, et al., 2006). Unfortunately, Dowker (2003) found that despite being an important skill for mathematical achievement, children are often poor estimators and estimation in the classroom typically consists of computing a formal algorithm and then rounding, rather than using approximation in meaningful ways.

The Problem. Despite efforts to reform education in the United States, children from East Asia outperform their American counterparts in mathematics achievement, and this is especially true for children from low-SES families (Ramani & Siegler, 2008; Sun & Zhang, 2010). Besides various school factors such as large class sizes, poorly prepared teachers, and fewer resources, some factors contributing to low performance occur well before a child enters school. Unfortunately, children from low SES backgrounds typically do not receive the same environmental exposure to early number activities, mathematical materials (such as board games), and language, and rarely catch up to their middle- and higher-SES counterparts in mathematical achievement (Ramani & Siegler, 2008; Levine, Suriyakham, Rowe, Huttenlocher, & Gunderson, 2010). This gap appears in children as young as preschool-age and widens with age (Clements & Sarama, 2007).

Low-SES children enter kindergarten behind middle-SES peers in various aspects of number sense, are less likely to utilize an accurate linear representation of number, which is a strong predictor of later achievement in mathematics, and show little growth in number sense into the beginning of first grade (Baroody, Lai, & Mix, 2006). Starting at such a disadvantage makes it extremely challenging to learn formal mathematics procedures or algorithms, or master more-advanced content like number combinations and multi-digit operations. It is no surprise that by grade 2 or 3 many children (particularly those who are low-SES) dislike and even fear mathematics (McLeod, 1992).

The Use of Models and Manipulatives. A powerful way for children to learn about abstract mathematical ideas is through models and manipulatives embodying various kinds of representations. The thought is that models and manipulatives- objects that can be seen, touched, moved, or felt—can serve as a concrete representation of an abstract concept, which in turn helps students to create mental representations of mathematical ideas and procedures (Mix, 2009). Mix (2010) discusses several mechanisms through which manipulatives can enhance learning, including by generating action, by acting as a conceptual metaphor, by offloading intelligence, and/or by focusing attention. Base-ten blocks, Cuisinaire Rods, Unifix cubes, an abacus, and other physical manipulatives have been used to teach about symbolic number concepts such as the base-ten system (Mix, 2009).

However, research regarding the effectiveness of manipulatives shows that simply employing their use does not guarantee a child will gain deep understanding of the underlying concepts (Mix, 2010; Moyer-Packenham, Salkind, & Bolyard, 2008). Clearly, the fundamental question is not whether or not manipulatives are effective, but rather in determining the specific circumstances in which particular models and manipulatives can be used most effectively for certain purposes. Moreover, children need enough time with a particular manipulative or representation to fully understand the underlying mathematical content. New representations must be carefully introduced with adequate scaffolding to connect multiple representations in meaningful ways (Sarama & Clements, 2009; Mix, 2010).

The value of technology to foster learning. The world is experiencing a major shift in education and learning. Particularly, the technological tools that children have at their disposal with which to work, play, and learn are significantly different than the pencil and paper and physical manipulatives of years passed. Technology shows the promise of enabling children to learn in the ways that researchers have found most effective. The following is a sampling of how technology can incorporate best practices suggested by learning theory:

- Make difficult abstract concepts visual and concrete
- Allow for repetition
- Provide consistent and appropriate feedback
- Modify content for varying developmental needs
- Personalize content
- Allow for practice and failure (without the feeling of disappointing a teacher or parent)
- Provide opportunities for success.

- Be engaging, appealing, and motivating
- Assist educators in assessing student progress

With the plethora of technological tools available at increasingly lower cost, including hardware such as handheld-devices, tablets, Smart Boards, computers TVs, cameras, and gaming consoles, as well as software such as applications (apps), games, virtual manipulatives, and digital curricula it is no surprise that children's access to technology has increased, even in low-income schools (Warschauer & Matuchniak, 2010).

Virtual tools (such as manipulatives and models) can visually represent mathematical ideas and relationships (Mix, 2010). Indeed computer technology can be used to create virtual manipulatives that in some ways may be more powerful than their concrete counterparts. For example, children can transform 100 virtual blocks into one hundred-block with the click of a mouse rather than by physically connecting each block one-by-one, which could be cognitively or even physically challenging (read, boring). The computer allows the child to focus on the relevant concept of base-ten rather than becoming distracted with the physical act of connecting the blocks, and may provide adaptive feedback in real-time.

However, the wide variety of technology available has translated into a dangerously wide variety of quality and use (Ginsburg, Carpenter, Labrecque, & Pagar, in press; Warschauer, Knobel, & Stone, 2004; Warschauer & Matuchniak, 2010). Although access to technology is becoming more equitable, teachers in low SES schools use the software in narrow ways, such as for drill practice of math facts, rather than for inquiry projects or building conceptual understanding of a topic (Clements & Sarama, 2008; Warschauer & Matuchniak, 2010). Additionally, low SES schools lack resources— including specialized technology personnel or professional development-- to support the teachers' use of technology. Despite such issues, technology can enable children to learn in the ways that researchers have found they learn best (Anderson & Kirkorian, in preparation) when it is well designed (Ginsburg, et al., in press) and implemented.

MathemAntics

MathemAntics (MA) is a computer-based software program designed to supplement math curriculum and teach children from age three to grade three about various number concepts in a playful and antic way. MA is divided into seven learning environments that address a different topics: Enumeration, Equivalence, Early Addition and Subtraction, Multiplication and Division, Written Calculation, Negative Number, and Estimation, each of which includes several game-like learning activities with the goals of elaborating and mathematizing a child's everyday mathematics (Baroody, Lai, & Mix, 2006; Ginsburg et al., 2006; Sarama & Clements, 2009), promoting integration between the child's informal everyday math with formal math knowledge, meaningful synthesis between formal mathematical representations (Vygotsky, 1986), and encouraging conceptual understanding.

The development of high quality math educational technology requires knowledge of a coherent theory of mathematics cognition and learning trajectories (Clements & Sarama, 2007; Ginsburg, 1989; Ginsburg, Jamalian, & Creighan, 2013), pedagogy, mathematics content, technological affordances, and game development. However, designing based on a strong theoretical framework does not ensure the technology will be effective at promoting learning. Formative usability research, which has historical origins in radio and television production, is a critical step in the development process, and can provide initial insight into whether or not the child can use the technology as intended, as well as assist in answering critical questions about the child's engagement with and learning from the technology (Flagg, 1990; Ginsburg et al., in press). Formative research drives necessary revisions to the software and can assist researchers in planning for formal learning studies that utilize traditional empirical research methods to study the effectiveness of the technology. The present study describes a coherently designed activity within MA that previously underwent formative usability research testing. (Creighan, in preparation; Ginsburg, Jamalian, & Creighan, 2013). In the formative usability testing, two researchers worked with a small sample of first and second graders on various activities from the Estimation Environment. Researchers observed as the children interacted with the activities, asked clinical interview questions to probe into the child's thinking, and took detailed field notes related to the child's motivation, usability of the software, and initial evidence for learning. Findings from the formative research inspired revisions to the software, including implementing control over the score-keeping algorithm and adding additional data captured in the computer log file.

Key Features of MANL. Here, I will elaborate on five key features of the Number Line activity within MathemAntics (MANL): (1) Content, (2) Multiple attempts, feedback, scaffolding of the midpoint, (3) Placing a range on the line rather than a single point, and (4) A meaningful reward system

(1) Content. As discussed, although difficult to define and measure, number sense is an important topic for the intended age group. Many successful interventions, including board games, computer-based software, and number lines together with physical manipulatives aim at improving various aspects of number sense in young children (4-5 years old) (Dyson & Jordan, 2011; Saxe, Earnest, Sitabkhan, Halder, Lewis, & Zheng, 2010; Griffin, 2004; Wilson et al., 2006; Ramani & Siegler, 2008; Ramani, Siegler, & Hitti, 2012). My goal is to explore whether MathemAntics can be used to improve number sense in low-SES second-grade children.

The number line is a model of symbolic number, and MANL requires children to directly interact with and place targets on a number line, thus making the number line a powerful virtual manipulative. Through interacting with the number line in this way, children can learn about numerical magnitudes and their relationships without mastering formal mathematical procedures or algorithms.

(2) Multiple attempts and scaffolded feedback. Multiple attempts and scaffolded feedback, where incremental hints are given to guide the child to the correct answer, transform incorrect answers into additional opportunities for learning (Malone & Lepper, 1987; Gee, 2005). As previously discussed, Petitto (1990) found that the most advanced strategy children used during number line estimation was the midpoint strategy, meaning the child mentally partitioned the line in half and placed the estimate using the midpoint as a reference. To encourage this strategy and discourage counting-based strategies, we show the child the midpoint after an incorrect first attempt.

(3) Placing a range of acceptable approximation or margin of error on the line rather than a single point. The traditional number line estimation task has a child mark a single point on number line to indicate his answer. When designing MANL, I wanted to focus on the idea of approximation. Therefore, the child is given a range to place on the line, rather than a single point. This feature was also implemented to discourage counting-based strategies.

(4) Meaningful reward system tied to user defined range mode. To make sure the educational technology is engaging, designers may draw from game theory. Gee (2005) states that a meaningful reward system can be a powerful motivator. To integrate the reward system into the activity, I (a) created a mode of the game in which rather than being given a FR to place on the line, the user has control over the size of his range (User-Defined Range mode) and (b)

tied the reward system to the size of the user's range, such that a child who is correct with a range of 5 units will receive significantly more points than a child who is correct with a range of 15. Clearly, it is more difficult to be accurate with a smaller range, causing a trade-off for the user between potential points and likelihood of being correct. Essentially, the child has control over how large or small to set the range of acceptable accuracy.

Beyond being motivational, the user-defined range mode functions both as a research tool and as a learning experience that introduces other important advanced math concepts and constructs. For example, the user-defined range feature can be used to investigate the psychology of risk-taking in mathematics, which has been subject to gender differences (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998), and further emphasize the idea of estimation and approximation by allowing the child to demonstrate confidence in their estimates (Byrnes, Miller, & Schafer, 1999; Clifford, 1991; Foersterling, 1980). It also introduces the idea of approximation in a very deliberate way. Finally, with some additional instruction this feature can also introduce basic ideas about statistics such as confidence intervals and variance. Because of the importance of this feature, we chose to conduct a between-subjects manipulation of this feature with two different experimental conditions. All other key features were constant in both experimental conditions, described below.

Findings from formative usability research with MANL interestingly revealed that when using the User-Defined Range mode of the activity, some children would stretch the range the entire length of the number line in order to be correct. Because of the small number of participants and flexible nature of formative research, conclusive inferences cannot be drawn from this observation. However, despite still receiving feedback on each trial, it is reasonable to assume that children who engage in "stretching" will not achieve the same learning gains as children who do not stretch. Formative research also revealed that contrary to Petitto's (1990) progression from counting-based strategies to partitioning strategies like the midpoint strategy, some children always answered quickly (and often accurately), without any visible strategy. The present study will follow up on these findings empirically, examining the stretching phenomenon and the quick response strategy.

The Present Study. The aim of the present study was to investigate how MANL can promote number sense skills in second-graders based on pretest and posttest measures and based on microgenetic data collected during the child's computer sessions. Secondly, I wanted to investigate whether the User-Defined Range manipulation impacts number sense differently than children who used a FR.

Research Questions. With regards to the overall effectiveness of MANL,

(1) Do children who play MANL outperform the Reading comparison group on a paper and pencil number line estimation measure at posttest? (2) Do children who play MANL utilize better expressed strategies at posttest than the Reading comparison group at posttest? (3) Do MANL conditions improve on standardized math measures? (4) Do children improve on MANL over time? (5) What features in the computer log file can be used to predict accuracy within MANL? (6) Are the accuracy and strategy effects of the User-Defined Range different from those of using a Fixed Range?

Chapter 2

METHOD

Design

This study utilized a unique methodology combining aspects of mixed pretest-posttest designs (including a combination of between and within-subjects variables) and microgenetic research. For pretest and posttest analyses, condition served as a between-subjects factor, and gender and school were controlled for. Although gender differences are not a primary focus of this study, there is evidence to suggest boys have a small advantage on estimation tasks and engage in more mathematical risk-taking, making it important to control for (Fennema, et al., 1998).

Participants

Participants were recruited from three schools in New York City serving primarily students from low-SES backgrounds were recruited as part of a larger pilot study of achievement for *MathemAntics* by fliers sent home to parents. On the consent form, parents could indicate whether they gave consent for their child to participate in the study, whether they would allow the sessions to be video or audio-taped, and whether they permitted the videos to be used in future educational settings.

Of the seventy-one students who returned signed consent forms, five were dropped throughout the course of the study, one because she did not speak English, two due to excessive absences, and two due to technical issues with the computer software, resulting in a final sample of sixty-six second-graders, 34 females, ranging from 6.53 to 8.72 years old (M = 7.33, SD = .045). Table 1 illustrates demographic characteristics of the three schools. Ethnic and socio-economic demographics were not collected from individual participants.

School	% Free and Reduced Price Lunch	Total Enrollment	Participants	% ELL	% Asian	% Black	% Hispanic	% White
A	84.9	539	20	7.4	1	79	19	1
B	73.3	623	11	11.6	1	54	44	1
C	66.5	785	35	20.0	53	8	25	14

Table 1: Demographic information of the three schools.

It is worth noting that even though all schools serve similar socio-economic status populations as evidenced by the percent of students who qualify for free and reduced-price lunches, School C, which accounts for more than half the sample, was quite different in terms of ethnic demographics. Research has shown that there are cultural differences regarding beliefs about education and mathematics specifically, with Asian parents valuing the importance of early academic training in mathematics more than do parents from Euro-American backgrounds (Parmar, Harckness, & Super, 2004).

Participants within each school were randomly assigned to one of three conditions: MA-User-Defined Range (UDR), MA-FR, or Reading Comparison (*n* = 23, 25, and 18, respectively). The two MA conditions were identical except for whether the range the child placed on the number line was a fixed width or was able to be defined by the user.

Procedure

Researchers worked one-on-one with all participants for six sessions: pretest, four computer sessions, and posttest. All researchers were blind to the study goals and hypotheses, and a different set of researchers blind to condition conducted all pretest and posttest sessions. In general, the child had one session per week unless he was absent on a particular day, in which case the following week he may be seen twice. During the computer sessions, the children in the two MA conditions (MA UDR and MA Fixed) played either MANL or another MA activity on 13- or 15- inch MacBookPro or MacBookAir laptops resulting in two sessions of each activity as the researcher coded observed strategies. The order of activities was counterbalanced across all participants. For the present study, I was only interested in the data from the two MANL sessions.

The two MANL sessions each consisted of one demo and 15 scored trials (resulting in 30 total scored trials) with these instructions playing on the first three trials of each session: "Here is a number line marked from zero [zero simultaneously pulses] to 100 [100 simultaneously pulses]. Where should we place [target number]? Use the mouse or arrow keys, and then press the space bar to submit your answer." For all trials, the child could take up to two attempts to answer correctly and received audio and visual feedback on all attempts. For correct answers, the target number appeared in the correct position on the number line while the audio recited, "Great estimate!" and the child's points earned for that trial were added to his score. For incorrect first attempts, a small gray arrow preserved the child's incorrect answer while the audio recited, "Oops, try again, your estimate was too high/low." For incorrect second attempts, a small gray arrow preserved the child's incorrect answer, the target number appeared in the correct position on the line as a labeled tick mark while the audio recited, "Oops," and the child's points were added to his score. The child then pressed the "next" button to proceed to the next trial.

Target numbers on both sessions were 50, 79, 92, 21, 77, 24, 61, 91, 79, 42, 9, 59, 61, 86, and 42. Because one of the observed strategies researchers coded for was the use of a Landmark (described below), the order of the targets was constant for all participants. For example, if the target numbers were randomized across participants and one child saw 91 and then 92, they might be more likely to use the landmark strategy compared to a child who sees 24 then 92.

Although the algorithm for the point system was different for both of the *MathemAntics* conditions, pilot work revealed that mean scores were similar for each condition. For the MA FR condition, the child received 10 points for correct responses, and 5 for incorrect responses. For the MA UDR condition, the score was calculated as an area under a curve (Base Score * (range of the NL)/Score Modifier) * (1-(sigma divisor), where base score was 5, score modifier was 10, and sigma divisor was 10, see Figure 2 for details.

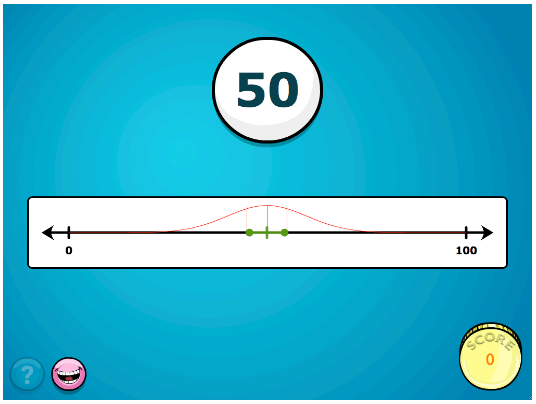

Figure 2: The score-keeping algorithm used in the User-Defined Range Condition. The area under the curve is the child's score, which decreases as the size of the range increases.

The reading comparison group, which controlled for working on a computer for 10-15 minutes with a researcher, played *Reader Rabbit: I Can Read with Phonics* (The Learning Company, 2006) reading software. Reader Rabbit has a similar format in that a child plays mini-activities with many trials. Because it is designed for first- and second-graders, we allowed the children to play up to three activities to ensure they stayed engaged for the same length of time as the MA groups.

Measures – Pretest/Posttest

Standardized Curriculum-based Measures (CBMs)

Whether the child saw the CBMs or the two Number Line tasks first was counterbalanced across all participants. Pretest/posttest sessions lasted approximately 15-20 minutes. Children were given four curriculum-based measures (CBMs) from mCLASS Math (Ginsburg, Pappas, Lee, & Chiong, 2011). The CBMs are two-minute timed paper-and-pencil measures used for screening at the beginning, middle, and end of the school year, or to monitor progress in key skills. They can be given frequently (weekly) and are designed to capture growth over time. Computation (solving multi-digit addition and subtraction problems) and Number Facts (one-digit addition and subtraction problems) are designed to tap into formal math knowledge, whereas Missing Number (completing a number pattern) and Quantity Discrimination (circling the larger of two numbers) are designed to measure number sense.

The CBMs come with standard instructions and prompts for each measure, and include a demo problem the researcher completes with the child. For Computation, the child receives a point for each correct digit in his answer, for Quantity Discrimination, the child receives a point for each correct answer and loses a point for each incorrect answer, and for Missing Number and Number Facts the child receives a point for each correct answer with no penalty for incorrect responses.

The CBMs were used to establish equivalence between conditions at pretest. To examine if prior number sense ability impacts the effectiveness of MANL, I calculated a categorical variable indicating whether or not the child scored above the median on the two number sense CBMs (Missing Number and Quantity Discrimination) to be used in data analyses. Additionally, I hypothesized that children in both MANL conditions would outperform the Reading comparison group on Missing Number and Quantity Discrimination at posttest. Since improved number sense could indirectly impact Computation and Number Facts, it is conceivable that children in the MANL groups could see gains in these measures as well.

Mental Number Line Task (MNL)

The MNL task was taken directly from items on the Test of Early Mathematics Ability (TEMA-3) (Ginsburg & Baroody, 1990), where the child is shown three numbers, and must verbally indicate which of the second two is closer to the first one, for example, “what number is closer to 4, 5, or 6?" In this adapted task, there were two demonstration items and six scored, and the child received a point for each correct response, thus receiving a score from 0-6. I hypothesized that children in both MathemAntics conditions would outperform the reading comparison group at posttest.

Number Line Estimation Task (NLE) - Accuracy

The Number Line Estimation Task (NLE) consisted of 1 demo and 9 scored trials where the child had to mark with a pencil where a target number belongs on a number line marked from 0-100 without feedback. All items were different target numbers than used in the computer sessions. To score this measure, I calculated a Percent Absolute Error (PAE) for each trial.

$$\text{PAE} = (\text{Actual Value} - \text{Observed Value}) / \text{Scale}$$

I calculated the Median PAE for each child at both pretest and posttest. Because MANL was designed based on this task, I expected that on average, children from both MathemAntics conditions would have a lower Median PAE than children from the reading comparison group.

Number Line Estimation (NLE) - Expressed Strategy

For the last two trials of the NLE measure, target numbers 36 and 57, the child was also asked for an expressed strategy and the researcher recorded the child’s response verbatim, resulting in two items at pretest and two items at posttest. I developed a coding scheme with a researcher who assisted with planning and data collection in the formative usability research, but did not assist with data collection of the current study. We began by examining the observed strategies coded for during the computer sessions (described below). Because in general, I was interested in counting-based strategies compared to number sense-based strategies, we started with four high-level categories: counting, number relationships, combination, and other. Then, using a random sample of fifty items (18% of the total number of items) from the study, developed sub-categories within the high level categories (see Table 2) that would also allow us to capture if the child described what they did and how they did it.

Category	Code	Definition	Example
Counting	Count-Simple	No detail about what was counted or how	"I counted"
	Count-Other	Detail provided, but not clear	"Counting the numbers"
	Count-Ones*	References counting by ones	"I counted 1, 2, 3, 4, all the way to 36"
	Count-Tens*	References counting by tens	"I counted 10, 20, 30"
	Count-Multistep*	Combination of counting by tens and ones	"I counted 10, 20, 30, and then 6 more to 36"
Number Relations	NR-Simple	No detail about what or how comparisons were used	"It's bigger"
	NR-Other	Detail provided, but not clear	"Because 3 is bigger than 2"
	NR-Sophisticated*	Mentions specific landmarks and how they were used	"I remembered 21 went here, so I moved up a little for 36"
	NR-Midpoint*	Used the midpoint as a landmark	"Because 50 is in the middle, so go up a little for 57"
Combination	Combo-Simple	Lacking clarity	"I counted again but it was too close to 100 so I moved it to the middle"
	Combo-Sophisticated*	Clearly states what landmarks were used and what was counted	"Because 20 goes here, then after 20 is 30, move up a little for 36"
Other	Other	Cannot be classified by any other code	"I just knew" / "I thought about it." / "It just goes there." / "I don't know."

Table 2: Coding scheme for NLE expressed strategy items. *Denotes explicit explanation

The additional researcher and I then used this coding scheme to code the sample of 50 expressed strategy items. The few discrepancies that arose were discussed and resolved. An independent coder blind to the study goals and condition and time point (pre or post) of all expressed strategies was trained on the strategy subcategories and given examples of each strategy. She then coded the sample of expressed strategies and had high internal consistency with the master coders ($K = 0.897$). She then coded all of the expressed strategies (including re-coding the sample).

Because MANL is designed to encourage number sense, I hypothesized children in both MathemAntics conditions would utilize fewer counting-based (and alternately more number relations) explanations at posttest compared to the reading control. Additionally, I hypothesized children in both MathemAntics conditions would be more explicit in their explanations than children in the reading control.

Based on the codes, I created three binary contrasts to test my hypotheses: whether or not the child utilized a counting-based explanation, number relations explanation, and explicit explanation (comparing the strategy codes labeled with an asterisk in Table 2 versus all other strategies) on at least one of the two items at pretest and posttest. An explicit explanation was defined as any strategy in which the child described what was done to solve the problem and provided details about how the problem was solved. It is important to emphasize this is not a measure of whether or not the child is using an advanced strategy, because clearly counting by ones to either of the target numbers (36 or 57) is less efficient than many of the other strategies. However, I felt that explaining, "I counted 1, 2, 3, 4, all the way to 36" indicated more proficient expression than a child who simply stated, "I counted."

Measures – Session Data (MA groups only)

For the MA groups, pretest and posttest scores were supplemented with session data consisting of computer log files, observed strategy for each trial, and Silverback (*Clearleft Inc.,* 2012) video recordings of each computer session, which is a screen capture software that simultaneously records (with parental consent) video and audio of the child with the laptop's built in web-cam. Computer log files captured data about the overall session such as child ID, session timestamp, activity played, and configuration settings used, as well as detailed information for each trial such as trial number, target number, trial percent absolute error (PAE), size of the range used in the user-defined range condition, and latency. The recording sheet was developed using strategies noted by Petitto (1990) and results from our own formative usability research. Table 3 defines all strategies coded by researchers. All session data allowed for a microgenetic analysis of important factors contributing to a child's performance (as measured by trial PAE).

Strategy	Definition
Count Forward	Counting up from zero
Count Backward	Counting back from 100
Count on	Counting from another number besides the endpoint
Midpoint	Clear evidence the child used the midpoint to partition the line and find the target number
Landmark	Child references another landmark (besides the midpoint) used to help find the target
Quick	Answers within 4 seconds
Other	No clear evidence of a strategy and answers after 4 seconds

Table 3: Definitions of observed strategies coded during computer sessions

Chapter 3

RESULTS

Pretest Analysis

I first established that conditions were equivalent at pretest using a MANOVA with Missing Number, Quantity Discrimination, Computation, Number Facts, Mental Number Line, and Number Line Estimation pretest scores as outcomes and condition as a factor, controlling for gender. Since children were blocked by school, it was not included in the pretest model. There were no significant differences between conditions, although I did see a significant effect of gender, Wilks' $\lambda = .791$, $F(5, 58) = 3.07$, $p = .015$. The univariate tests reveal that boys did significantly better than girls on Number Facts, $F(1,62) = 4.08$, $p = .048$, Quantity Discrimination, $F(1,62) = 11.87$, $p = .001$ and Computation, $F(1,62) = 6.51$, $p = 0.13$. Boys were marginally better than girls on Missing Number, $F(1,62) = 3.02$, $p = 0.087$. There were no differences on the Mental Number Line task.

	Female		Male	
	Mean	St. Dev.	Mean	St. Dev.
Number Facts	9.15	5.56	12.16	7.00
Missing Number	4.29	2.84	5.34	2.99
Quantity Discrimination	4.91	4.74	8.81	4.53
Computation	4.74	3.26	7.66	6.40

Table 4: Gender differences in scores on Curriculum Based Measures at pretest

To determine that all three conditions were equivalent on the expressed strategy measure, I conducted chi-squared analyses on three pretest binary variables: counting, number relations, and explicit explanations. All three were non-significant, indicating that all three conditions were similar at pretest, χ^2 (1, N = 66) = 0.24, p = 0.89, χ^2 (1, N = 66) = 0.35, p = 0.84, χ^2 (1, N = 66) = 1.64, p = 0.44, respectively.

Question 1: Do both MA conditions outperform the Reading comparison condition at posttest on the Number Line Estimation (NLE) measure?

Because I did not block by pretest, I first ensured all three conditions had roughly equivalent numbers of participants who scored below the median on the two Number Sense CBMs (Quantity Discrimination and Missing Number), so that it could be included in the analysis as a factor. Indeed, chi-square tests revealed no significant effect of condition.

An Analysis of Covariance using posttest Median PAE as the outcome, condition and whether the child scored above or below the median on the two number sense CBMs (Quantity Discrimination and Missing Number) as factors, controlling for gender and school location (Chinatown or Harlem), and pretest NLE score as covariates was conducted starting with the full factorial model and utilizing stepwise removal of non-significant interactions. A significant interaction between Condition and Number Sense ability remained significant in the final model, $F(2, 56) = 6.44$, $p = 0.003$ (see Figure 3). Post hoc analyses using the Tukey's HSD procedure revealed that when children scored below the median on the number sense CBMs at pretest, they did worse in the MA UDR condition than the MA FR condition as evidenced by a higher average Median PAE at posttest $t(56) = 2.93$, $p = .005$. Conversely, when children scored above the median on the number sense CBMs at pretest, they did significantly better in the MA UDR condition than the MA FR condition at posttest $t(56) = -2.12$, $p = .038$[1]. This seems to indicate that children with low number sense ability at pretest benefit more from the MA FR condition rather than the MA UDR condition. Conversely, children with high number sense ability at pretest benefit more from the MA UDR condition than the MA FR condition.

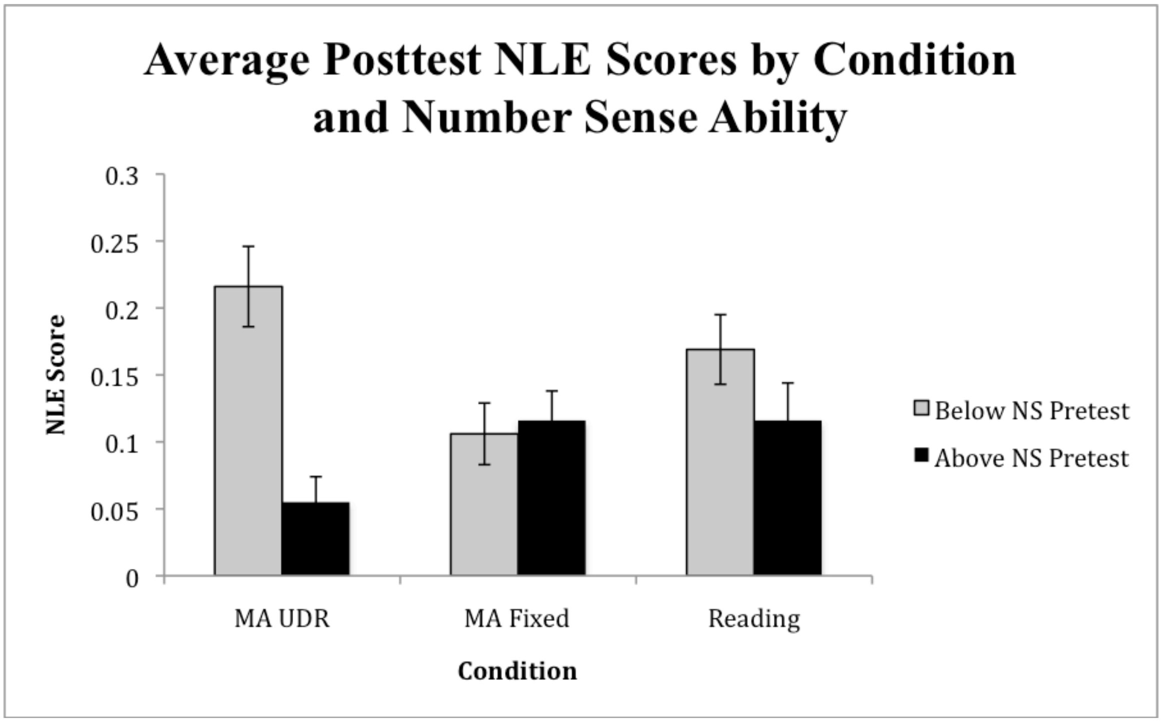

Figure 3: Average scores on the Number Line Estimation posttest by condition and pretest number sense ability.

Question 2: Do both MA conditions outperform the Reading comparison group on expressed strategy items on the Number Line Estimation posttest?

A Chi-Squared analysis comparing whether or not the child used a Counting-based expressed strategy on at least one of the two posttest items revealed a significant effect of condition, χ^2 (2, N = 66) = 10.247, p = .006. Table 5 shows that children in the MA UDR condition are using counting-based expressed strategies much less than expected by chance, whereas children in the Reading comparison condition are using counting-based expressed strategies more frequently than expected. In the MA UDR condition, of the 15 children who counted at least once at pretest, four exhibited no counting at posttest (a 26% decrease) and no children who did not count at pretest counted at posttest.

	MA UDR (n = 23)		MA Fixed (n = 25)		Reading (n = 18)	
	Observed	Expected	Observed	Expected	Observed	Expected
No	23	18.1	18	19.7	11	14.2
Yes	0	4.9	7	5.3	7	3.8

Table 5: Observed versus expected counts of using Counting-based explanations on at least one of the two posttest expressed strategy items.

A chi-squared analysis comparing whether or not the child used a Number Relations expressed strategy at least once at posttest revealed a marginally significant effect of condition, χ^2 (2, N = 66) = 5.83, p = .054. Looking at the observed versus expected counts shows MA UDR counting significantly less than expected by chance and reading counting more than expected, but no effect within MA Fixed (See Table 6). Within the MA UDR condition, 11 children who did not use Number Relations strategies at pretest used Number Relations strategies at least once at posttest (a 47% increase) compared to only four children in the Reading Comparison group (a 22% increase).[2]

	MA UDR (n = 23)		MA Fixed (n = 25)		Reading (n = 18)	
	Observed	Expected	Observed	Expected	Observed	Expected
No	8	12.2	14	13.3	13	9.5
Yes	15	10.8	11	11.7	5	8.5

Table 6: Observed versus expected counts of using Number Relations on at least one of the two posttest expressed strategy items.

Therefore, we have evidence to suggest that children in MA UDR are counting less and using number relations more than the reading comparison in their expressed strategies at posttest.

Similarly, there is a significant effect of condition, χ^2 (2, N = 66) = 7.39, p = .025 for whether or not children utilized an explicit explanation at least once at posttest. Looking at the observed versus expected counts, MA UDR results in using explicit explanations more frequently than expected, whereas the reading comparison is using explicit explanations less frequently than expected (see Table 7). Within the MA UDR condition, 14 children went from not using an explicit explanation at pretest to using an explicit explanation at least once at posttest, or a 60% increase, compared to only five children in the Reading comparison group, or a 27% increase. Children in the MA UDR condition are using more explicit explanations at posttest than children in the Reading comparison condition. A simple explanation for the differences between the MA UDR condition and the MA FR condition could be that the extra step in the MA UDR condition provides children with more time on task. To test this, I conducted an analysis of variance using trial time as the outcome, condition, session, and question as factors, student ID as a random factor, and controlling for gender and school location revealed no significant effect of condition, $F(1, 1{,}408) = 1.05$, $p = 0.31$.

	MA UDR (n = 23)		MA Fixed (n = 25)		Reading (n = 18)	
	Observed	Expected	Observed	Expected	Observed	Expected
No	7	11.8	14	12.9	13	9.3
Yes	16	11.2	11	12.1	5	8.7

Table 7: Observed versus expected counts of using an explicit explanation on at least one of the two posttest expressed strategy items.

Question 3: Do children in the two MA conditions outperform the reading comparison condition on mCLASS CBMs and the MNL task at posttest?

To examine whether MA positively impacted CBMs or MNL, I ran a MANCOVA with Missing Number, Quantity Discrimination, Computation, Number Facts, and Mental Number Line posttest scores as outcomes and condition, gender, and school location (Chinatown or Harlem) as factors. There were no significant effects of condition or gender, but there was a significant main effect of school location, Wilks' $\lambda = .70$, $F(5, 52) = 4.46$, $p = .002$. Univariate tests show that on average, children from the Chinatown school are outperforming children from the Harlem schools on Number Facts, $F(1, 56) = 12.51$, $p = 0.001$, and marginally on Quantity Discrimination, $F(1, 56) = 3.92$, $p = 0.056$. Pretest gender differences are no longer significant at posttest. Therefore, there is no evidence to suggest that condition impacted scores on any of the CBMs or the MNL task.

Question 4: Do children improve on MANL over time?

Using the computer log files, I calculated a Median PAE score for each child for MANL Session 1 and MANL Session 2. I conducted a repeated measures Analysis of Covariance using Median PAE as the outcome, Time (MANL Session one or two) as a within-subjects factor, Condition as a between-subjects factor, controlled for Gender and School location, and added Missing Number and Quantity Discrimination pretest scores as covariates to determine whether median PAE decreased from Session 1 to Session 2. Regardless of other factors, there was a significant effect of time, $F(1, 44) = 12.83$, $p = 0.001$ (see Figure 4). Interestingly, pretest NLE score is not predictive of Median PAE at Session 1. Similarly, Median PAE at session 2 is not predictive of posttest NLE score. In general, PAEs are much lower during the computer session than on the pretest and posttest, which may be a result of receiving feedback during the computer sessions.

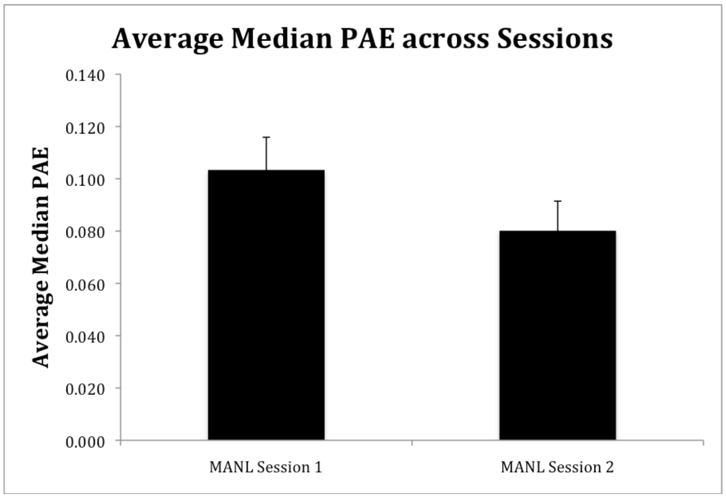

Figure 4. Median Percent Absolute Error by session.

To further answer this question, I investigated how observed strategies changed from session 1 to session 2. Children's observed strategies consisted primarily of quick and other, with very few instances of counting, midpoint, or landmark strategies. Counting occurred only on 4.96% of trials, but a Chi-squared analysis showed that by Session 2, children were counting less frequently than expected, χ^2 (1, N = 719) = 9.56, p = 0.002. Based on the chi-squared distribution, it is expected that children will count on 15.5 trials at session one, and 15.5 trials at session 2, but the observed data show children count on 24 trials at session one and 7 trials at session 2.

At session 1, children in the MA UDR group are answering quickly less frequently than children in the MA Fixed group, but that difference is eliminated by session 2. To test this empirically, I ran a I conducted a Generalized Linear Mixed Effects Model with whether or not the child was coded as answering Quick as the outcome variable and with Student ID as a random factor, condition as a between-subjects fixed factor, and session as a within-subjects fixed factor. I controlled for gender and school differences, beginning with a full factorial model and stepwise removing non-significant interactions. An interaction between condition and session remained in the model, $F(1, 1{,}413) = 5.377$, $p = .012$. Post hoc analyses using the Bonferroni correction reveal that MA UDR participants answer more quickly in session 2 than in session 1 t (1,423) = 5.69, adjusted $p < .0001$. Indeed, the data supports my hypothesis that children would improve on MANL over time.

Question 5: What factors of the computer log file and observed strategies predict trial accuracy?

To determine which features best predict trial Percent Absolute Error (PAE), I conducted a Generalized Linear Mixed Effects Model with trial PAE as the outcome variable and with Student ID as a random factor, condition as a between-subjects fixed factor, and session, trial number, target number, and whether or not the researcher coded them as answering quickly as within-subjects fixed factors. I controlled for gender and school differences, beginning with a full factorial model and stepwise removing non-significant interactions. For significance tests, I conducted pairwise contrasts using the Bonferroni correction.

The final model included a significant interaction between target number and whether or not the child answered quickly on that trial, $F(11, 1{,}412) = 2.49$, $p = 0.004$, (see Figure 5) and a significant main effect for MANL Session, $F(1, 1{,}412) = 14.873$, $p < 0.001$. A post hoc analysis using the Bonferroni correction revealed that when the trial was coded as a quick response, PAE was significantly higher than when the trial was not coded as a quick response for target numbers 21, 79, and 91, $t(1{,}409) = -2.09$, 3.31, and 4.08, $p = 0.037$, 0.001, and 0.000, respectively, and marginally significantly lower when the target number was 92, t (1,409) = -1.816, adjusted $p = 0.07$. When a trial is coded as a quick response, target number does not impact trial PAE, but conversely, when a trial is not coded as a quick response, target number significantly impacts trial PAE. Out of the 66 post-hoc comparisons, Table 8 shows only the twelve significant differences for ease of interpretation. In general, if a trial is not coded as quick, trial PAE is more likely to be higher when the target number is below 21 and above 79.

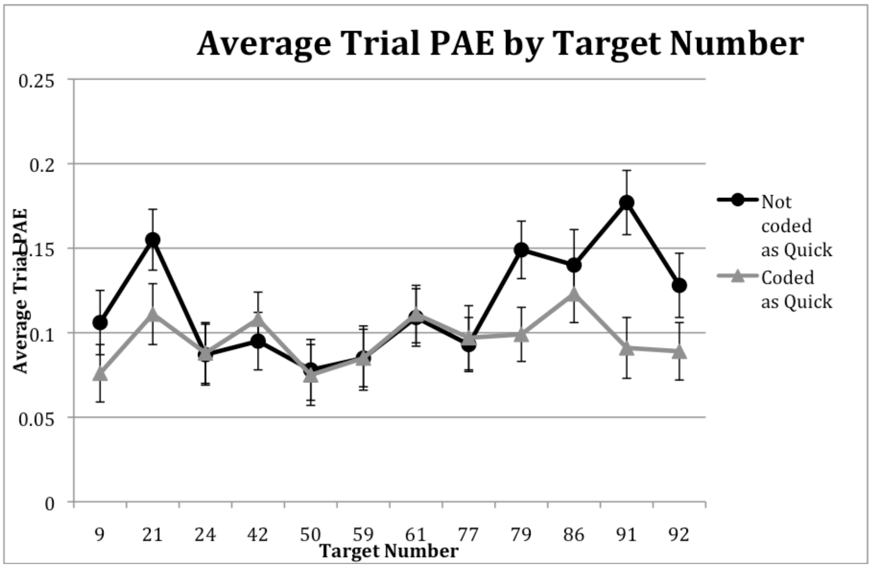

Figure 5: Average Trial PAE by Target number. Target number impacts PAE when a child is not coded as answering quickly.

Pairwise Comparison Target Number	*t*	df	adj. *p*
21-24	3.36	1,409	0.04
21-42	3.59	1,409	0.02
21-50	3.79	1,409	0.01
24-79	-3.59	1,409	0.02
24-91	-4.28	1,409	0.00
42-79	-3.94	1,409	0.01
42-91	-4.57	1,409	0.00
50-79	-4.07	1,409	0.00
50-91	-4.69	1,409	0.00
59-79	-3.45	1,409	0.03
59-91	-4.15	1,409	0.00
77-91	-3.89	1,409	0.01

Table 8: For trials not coded as quick: significant differences between trial PAE based on target number

A post hoc analysis using the Bonferroni correction reveals that trial PAE is lower for session two than session one, $t(1,412) = 3.857$, adjusted $p < .001$. Above and beyond other factors captured from the computer log file, whether or not the child was coded as answering quickly for a given trial significantly impacts Trial PAE. Further, when a trial is not coded as a quick response, target number plays an important role in predicting Trial PAE. Specifically, children who do not answer quickly tend to have a higher Trial PAE when the target number is 21, 79, 91, and 92.

Question 6: Are the effects of MANL different for MA UDR than MA FR condition?

Because the User Defined Range is a key feature of MANL, I used the additional data from the log files about the size of the range the child used on each trial to conduct analyses for the MA UDR condition only. I began with an exploratory analysis examining what children did with the UDR. Figure 6 shows a box plot of each child's ranges used across all 30 trials of MANL. From this figure, it is clear that there are two children who stretched the range the entire length of the number line for the majority of trials. Additionally, there are three children (Student ID 305, 351, and 382) who never manipulated the range at all and left it the default size of ten units, and several other children who kept it the default size for the majority of trials, as their box plots are represented by a line or a line with a few outliers.

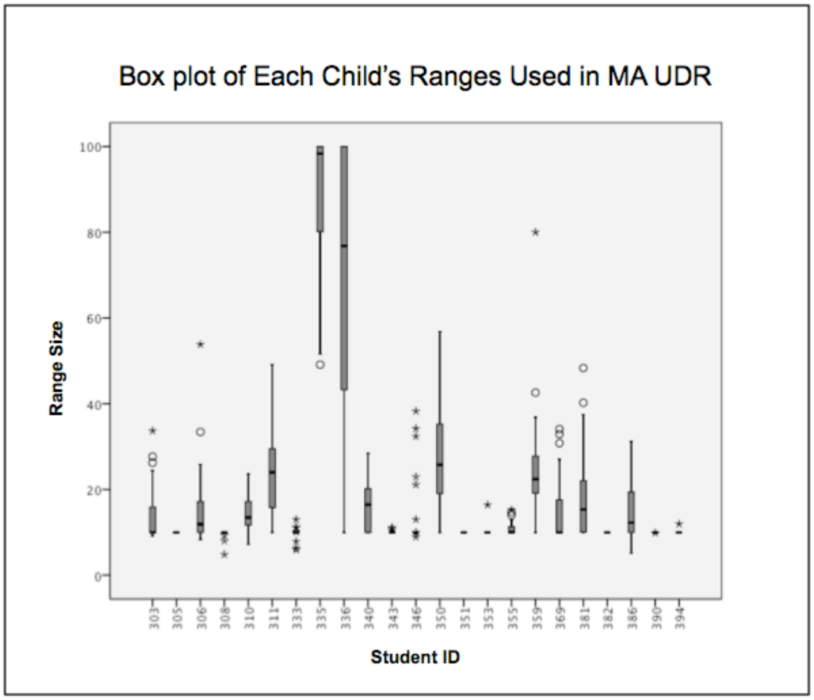

Figure 6: Box plots of each child's ranges used within the MA UDR condition.

I calculated each child's average range size and variance in range sizes used across all 30 trials and correlated it with their Median Computer Session PAE to investigate the relationship. After removing the two stretchers and three children who never manipulated the range, Median Computer Session PAE is moderately correlated with average range size $r(16) = 0.42$, $p = 0.08$, and is significantly correlated with variance in ranges used $r\ (16) = .51$, $p = 0.03$. This indicates that variance in ranges used is slightly more predictive of performance than average range size used,. Thus, a child who is consistent in the size of ranges used across trials (low variance) is more likely to do better in the activity (low PAE). Further, a child who is inconsistent with the size of ranges used (high variance) is more likely to have a higher average range, $r\ (16) = .95$, $p < 0.001$, with stretchers and default-only children excluded, $r\ (21) = .72$, $p < 0.001$, with the full sample.

Next, I examined whether or not number sense ability as measured by the sum of scores on Missing Number and Quantity Discrimination measures at pretest predicted the child's variance in ranges used within MA UDR. A regression using variance in ranges used as the outcome, pretest number sense score as a numeric predictor revealed that when controlling for gender and school location, pretest number sense score significantly predicts variance in ranges used t (21) = -2.91, p = 0.008, such that the higher the child's number sense score at pretest, the more likely the child is consistent with ranges used (lower variance).

Finally, to examine whether range size affects trial PAE beyond additional factors from the GLM model using log files from both MA conditions, I conducted a regression analysis with trial PAE as the outcome, range size, initial number line click value, and initial number line click time[3] as continuous predictors, and MANL Session as a categorical predictor. Because the target number significantly impacted trial PAE in the previous model with children doing particularly worse when the target number was 21 and below and 79 and above, I also included whether or not the question was less than 21, and whether or not the question was greater than 79 as categorical predictors, controlling for school location and gender, and including Student ID as a random factor. The omnibus test for the analysis was significant, $F(9, 671) = 71.56$, $p < 0.001$, with whether or not the question was greater than 79 (Question = High), Range Size, Initial number line click value, initial number line click time, and MANL Session all significantly predicting trial PAE (see Table 9).

Variable	B	St. Error	*ß*	*t*	Sig. (*p*)
(Constant)	0.15	0.05		3.16	0.00
School Location	0.00	0.01	-0.01	-0.31	0.76
Question = Low	-0.01	0.01	-0.02	-0.61	0.54
Question = High	0.06	0.01	0.22	6.63	0.00
Range Size	0.00	0.00	0.59	18.97	0.00
Initial Click Value	0.00	0.00	-0.26	-6.55	0.00
Initial Click Time	0.00	0.00	0.08	2.72	0.01

MANL Session	-0.02	0.01	-0.06	-2.26	0.02
Gender	0.01	0.01	0.02	0.83	0.41
Student ID	0.00	0.00	-0.03	-1.06	0.29

Table 9: Regression model predicting Trial PAE using factors from the computer log files.

Using the additional data of the child's range size, I was able to further explore the relationship between number sense and MANL within the MA UDR condition. Specifically, the data shows that children with low number sense abilities at pretest are more likely to utilize larger ranges on average and less consistent ranges from trial to trial, and have higher Trial PAEs. Other features from the log file that significantly predicted Trial PAE when looking across both MA conditions remain important when looking specifically within the MA UDR condition, like how quickly a child is answering, MANL session number, and when the target number is equal to or greater than 79.

Chapter 4

DISCUSSION

Overview of Findings

In general, this study examined the success of MANL in improving second-graders' number sense.

(1) Number sense ability at pretest impacts the effectiveness of condition

Children with low number sense ability at pretest (as measured by scoring below the median on Quantity Discrimination and Missing Number) did significantly worse on the paper-and-pencil number line estimation posttest measure when they were in the MA UDR condition than the MA FR condition. On the contrary, children with high number sense ability at pretest did significantly better on the paper-and-pencil number line estimation posttest measure when they were in the MA UDR condition than the MA FR condition. The hypothesis that children in both MA groups would outperform the reading comparison group at posttest on the paper-and-pencil number line estimation measure is not confirmed. Rather, it seems as though prior number sense ability impacts the effectiveness of the MA conditions.

When children with low number sense do not have the freedom to set their own range as in the MA FR condition, they outperform children with low number sense in the MA UDR condition on the paper and pencil posttest. Although this result was unanticipated, the computer log files from both MA conditions allowed me to investigate this further. As discussed in more detail below, children with low number sense ability at pretest are less likely to utilize the MA UDR feature efficiently, as evidenced by their use of larger and more varied ranges. By contrast,, children with high number sense ability performed better on the number line estimation posttest when they were in the MA UDR condition than the MA FR condition. A possible explanation is that the MA FR condition was not challenging enough for children who already have strong number sense abilities. Research has shown that the appropriate level of challenge promotes learning and improved performance (Gee, 2005). The extra step in the MA UDR condition may require the child to think more deeply about estimation or reflect on how confident he is in his estimate, thus making the child perceive the task as more challenging. However, more research is needed to investigate this claim more thoroughly.

(2) Children in the MA UDR condition used better expressed strategies at posttest

Another major finding is that children in the MA UDR condition (but not the MA FR condition) used fewer counting-based expressed strategies, more number relations expressed strategies, and more explicit explanations at posttest than the Reading comparison group. This is promising, although unanticipated because I expected both MA conditions to see the benefits. Why does the effect only hold for the MA UDR condition? A simple explanation could be that the extra step in MA UDR gave these children significantly longer time on task. However, trial average trial time does not differ by condition. Another possible explanation is that the ability to manipulate the range in MA UDR increases metacognitive awareness for the activity, and metacognitive awareness is required for the expressed strategy measure. Although this explanation is plausible, more research is needed to investigate the role of metacognition in the expressed strategy measure, or other potential mechanisms through which MA UDR is positively affecting children’s expressed strategies.

(3) Children in the MA conditions did not outperform the Reading comparison on standard measures

Although I anticipated children in both of the MA conditions to outperform children in the Reading comparison condition on the two Number Sense Curriculum-Based Measures and the Mental Number Line task at posttest, the data from this study does not support that claim. Similarly, there were no differences on the CBMs of formal math abilities (Computation and Number Facts). There are several explanations to account for this result. Firstly, these children only had four computer sessions lasting 10-12 minutes, or less than 45 total minutes of intervention. It is reasonable to speculate that the limited dosage was unable to influence performance on these measures above and beyond the child's standard math curriculum. Future research is needed to investigate the effects of playing MANL regularly over a longer time span.

The Mental Number Line task only contained six items each of which the child had a 50% chance of getting the correct answer, and thus we encountered a ceiling effect. An improved measure with more items that also captures response time should be used in future research as a better measure of number sense. Future studies may also consider incorporating different measures of Number Sense than the Quantity Discrimination and Missing Number CBM. Both of these measures involve symbolic number, which may in fact be measuring number sense access rather than pure number sense, and some problems involve mathematical symbols (e.g., in Quantity Discrimination, choosing between 50+2 and 60+2). Although strong number sense may aid a child in realizing the +2 is the same in both problems and thus can be ignored, many children saw the math symbols and immediately began solving the algorithms by a learned procedure. It is difficult to attribute this to poor number sense alone, as children may have thought that was how the problems "had" to be solved. Incorporating measures of non-symbolic quantity discrimination, non-symbolic estimation (where a child must quickly estimate the number of dots on the screen), or non-symbolic to symbolic mapping (where a child is given a non-symbolic representation of dots on the screen and must choose between two symbolic numerals that matches the quantity or vice versa) may help further unpack the relationship between MANL and number sense (Brankaer, Ghesquière, & De Smedt, 2014).

(4) Children improve on MANL over time

In general, children improve on MANL over time, regardless of condition, as evidenced by Average Median PAE decreasing from session 1 to session 2. This is not surprising, as repeated practice with an activity with feedback typically leads to improved performance over time (Gee, 2005). Although Median PAE is typically used for the number line estimation measure, the computer log files allowed me to examine children's performance over time microgenetically.

In general, the most frequently coded strategies were Quick and Other, with very few instances of Counting, Midpoint, and Landmark strategies. In contrast to Petitto's (1990) findings, I did not see children progress from counting-based strategies to partitioning-based strategies like the midpoint strategy. This could be due to the children in this study being older than her sample and thus already surpassing her proposed trajectory of strategy development, or suggest that some children answer quickly without needing to utilize a strategy due to strong number sense or number sense access. Alternatively, it is possible that children in this study did begin to utilize partitioning strategies more often over time, but a lack of clear behavioral evidence that the child was using such strategies could have resulted in the trial being coded as Quick or Other.

Despite seeing few instances of counting across both sessions, by their second session children are abandoning inefficient counting strategies by session 2, providing evidence they are improving on MANL over time. It is not surprising that the extra step in the MA UDR condition requires makes it less likely they will be coded as having a quick response. It is interesting to note that by session 2 these children begin answering quickly just as often as children in the MA Fixed condition. Children in this condition may just be getting faster at setting the size of their range, or may less likely to manipulate the range at all during session 2. Although not a goal of the current study, future analyses using this data set could attempt to disambiguate this finding by examining whether or not use of the default range size significantly increases from session 1 to session 2.

(5) Factors from session data can predict trial PAE

When looking at the log files and observed strategy data from both MA conditions to see which factors best predict trial PAE, whether or not a response is coded as quick determines whether or not target number significantly impacts trial PAE more so than other factors such as trial number. NL session is another significant predictor, with trial PAE being significantly lower for session 2 than session 1, providing further evidence to support improvements over time.

It is not surprising that overall, when children answer quickly they are more likely to have a lower PAE than when they are if not coded as answering quickly. A reasonable explanation is that children who answer quickly are utilizing their strong number sense or access to it while playing MANL. Children without strong number sense may be resorting to alternative mechanisms or strategies like counting to find the target number, potentially explaining why when a child does not answer quickly, target number is an important predictor of trial PAE. The fact that when children do not answer quickly, they are more likely to have a higher PAE when the target number is high (=>79) is consistent with the literature that suggests these children may be utilizing a logarithmic representation rather than a linear representation of number, and are thus devoting more space on the number line for smaller target numbers and placing larger target numbers close together. To investigate individual differences in the number line estimation task, researchers plot a participant’s actual responses against the target number and plot a best-fit function. Perfect performance on the number line estimation task (no error on any trial) would fit the linear function x = y, meaning the participant marked the target number in its exact position. Therefore it is optimal that a linear function fit the data. However, kindergartners’ and first-graders’ plots typically fit a logarithmic function (see Figure 1). Although by second grade most children are utilizing a linear representation of number, a similar method could investigate whether children who typically do not answer quickly are still utilizing a logarithmic representation of number.

However, it is interesting to note that when not answering quickly, children were also more likely to have a higher PAE when the target number was low (<=21). This suggests that something besides a logarithmic representation of number may be accounting for this finding. It could be that when a child does not answer quickly for these lower target numbers, he is resorting to counting from zero to find the target number. Although not a goal of this study, analyses investigating whether trials in which the researcher coded the child as using a counting strategy occur more frequently with specific target numbers could shed further light onto this finding. Future studies using improved methods of determining whether or not a child is counting could also investigate these claims.

It is also interesting to note that trial PAE was significantly lower when the target number was 24 than when it was 21, despite the two numbers being only three units apart. It is possible that since it is framed as an estimation task, children weight the tens and ones digits differently for different target numbers. For example, 21 is close to a decade number, and thus the child may be ignoring the ones digit altogether and aiming for “somewhere close to 20.” Whereas when the target number is 24, the ones digit plays a more important role in determining the target number's position in relation between the decades. As this was not the primary goal of this study, I did not look at how the data could support this claim. However, using the current data set, one could investigate whether average trial time is higher for target numbers that fall between decades than target numbers close to decades. However, future research using a wide variety of target numbers is needed to provide more conclusive evidence for this claim.

(6) Data from the ranges children use within the MA UDR condition can predict trial PAE

The most interesting findings come from taking a closer look within the MA UDR condition. There were clear individual differences on what children did with the ranges, with two children engaging in "stretching" on the majority of trials, three children leaving the range the default size on every trial, and several others leaving the range the default on many trials. Indeed, the data revealed that children who scored lower on the two number sense CBMs at pretest were more likely to be inconsistent with the sizes of ranges used from trial to trial, utilized larger ranges, and had higher trial PAEs.

This, coupled with the previous finding of pretest number sense ability impacting the effect of condition on posttest number line estimation scores, indicates that children with low number sense are not using the user-defined range feature as an effective learning tool, and are thus not seeing significant learning gains over time. Conversely, children with high number sense at pretest saw significant learning gains in the MA UDR condition over children with high number sense in the MA FR condition.

Future research is needed determine why children with low number sense are using the tool ineffectively. It is possible that these children are more likely to have low math self-efficacy or it could be explained by a motivational construct like goal orientation, the purposes for engaging in achievement behavior and signify a standard by which the individual judges his/her performance and attributes the cause of success or a failure of an outcome (Pintrich, 2003), or mindset (Dweck & Legget, 1998). Children with low number sense may be utilizing a performance goal orientation, or making the range large to ensure they will get the answer "right," whereas children with high number sense may be more likely to utilize a mastery goal orientation or mastery mindset and try to challenge themselves by using smaller ranges at the expense of getting the answer correct. Alternately, children with low number sense may not be using the UDR as a learning tool, but simply playing with adjusting the range as a fun feature. Unfortunately, nothing within the current data can provide insight into this issue, and thus additional data with measures of these and potentially other constructs are needed.

Implications

The findings from this study have implications for researchers, educational technology designers, and math educators, which are discussed in detail.

The findings from this study are consistent with other number line estimation research that shows children improve with the task over time with repeated practice and feedback and research that suggests number sense is subject to individual differences (Hyde, Khanum, & Spelke, 2014; Jordan et al., 2006). Although we found promising evidence of MANL improving number sense as measured by improved performance on MANL over time and results from the number line estimation measure of accuracy and expressed strategy, data from this study did not clearly explicate the extent to which second-grader's number sense abilities are malleable. In addition to the avenues of future research discussed above, a necessary next step for researchers will be to further investigate the mechanisms through which MANL impacts number sense, and effects of children using MANL over a longer period of time.

The findings from this study also have important implications for the design of educational technology. MANL was designed based on cognitive psychology literature on the development of the understanding of number and number sense, theories of children's learning in general, and game theory. However, it is interesting to note that often children in the MA FR condition did not see the same learning benefits as children, particularly with high number sense, in the MA UDR condition. This seems to indicate that providing feedback and multiple attempts alone is not always sufficient in transforming a task used in psychological research into a meaningful learning activity. Although children with low number sense benefited more from the MA FR condition as evidenced by lower Median PAE on the number line estimation posttest than did low number sense children in the MA UDR condition, children with high number sense did not benefit from the MA FR condition as measured by the same number line estimation task. In addition, children in the MA FR condition did not see the same benefits on the expressed strategy measure at posttest as did children in the MA UDR condition. This suggests that the UDR feature adds something extra to the task that enhances learning, whether it is by increasing metacognitive awareness, increasing the perceived level of challenge for the activity, or through an alternative mechanism. Many activities based on number line estimation have been developed, often times requiring the child to complete multiple trials of the task situated within a narrative. For example, a fisherman in a boat is trying to catch a fish, and the user must put the fisherman's boat in the correct position to find the fish. It may be that repeated trials with feedback within a narrative version of the activity is enough to promote learning above and beyond a version of the activity lacking narrative.

However, the implementation of the UDR feature added an extra step, which when used effectively shows positive effects on learning beyond using MA FR, despite both conditions sharing all other key features such as feedback and multiple attempts. Although I hypothesize that this extra step is encouraging the child to think about the task more deeply and deal with the concept of approximation, risk taking, and confidence in estimates very explicitly, more research is needed to explore the mechanisms through which MA UDR is positively affecting performance within the computer sessions and long term effects on other mathematical abilities. Future research may also investigate how manipulating various aspects of MA UDR impact the way in which children interact with the tool. For example, the score-keeping algorithm could be changed so as to punish a child for an incorrect response or for stretching the UDR too large. For the present study, I did not want to subject any possible poor estimators to negative scores. Alternatively, I set the default range size of the MA UDR condition to be the same size as the MA FR condition. However, there were very few instances of children making the UDR smaller than the default. Future studies could examine how varying the default range size impacts children's behavior in interacting with it.

This study also demonstrated how well designed educational software provides meaningful data to researchers. Although this study showed which factors from the computer log file and observed strategies successfully predict Trial PAE, there are many additional analyses that can be conducted utilizing this data. For example, this study only examined data from the child's first attempt. Different factors may be important in predicting trial PAE for the second attempt. There may be different learning implications for children who require both attempts on the majority of trials. For example, a child who is consistently incorrect on the first attempt gains scaffolded feedback and more practice on the activity than children who are typically correct on the first attempt. The child's incorrect answer is preserved on the number line and the midpoint is labeled, which may not only provide scaffolding for the current trial, but may impact the child’s response to subsequent trials. Although outside the realm of the present study, future analyses could examine the impact of utilizing two attempts on performance on subsequent trials. This highlights how looking at the data microgenetically at the trial level provides much more detailed and useful information than simply looking at the child’s median PAE at each session.

The current study may serve as a model for how researchers and game designers alike should approach the evaluation of educational technologies. MANL was developed based on sound principles from cognitive psychology and game design, and underwent several rounds of formative evaluation before the present study. Although I hypothesized that both the MA UDR and the MA FR conditions would see learning gains, results showed MA UDR to be much more promising for promoting learning. By designing the study in an empirical way to carefully compare the two groups, I saw explicitly that simply manipulating one feature of an activity can have powerful effects on learning.

Just as designers can benefit from incorporating psychological research methods to evaluate their designs, researchers can learn from the benefits of conducting research in real world settings instead of laboratories. Although this study benefited from having researchers work one-on-one with the child, something unlikely to happen in a classroom setting, I dealt with many of the challenges that face schools today. Several of the classrooms had large class-sizes (>25), making teachers sacrifice instruction time to classroom management. Children coming from classrooms lacking structure were often rowdy, requiring the researcher to attempt to manage the child’s behavior during the sessions. All three schools lacked a completely quiet workspace for our data collection, leading to us work with children in libraries being shared with several other groups or cafeterias for small windows of time between breakfast and lunch service. And finally, Hurricane Sandy occurred three weeks into data collection, causing all three schools to close for a week, one of which reopened without heat. Although the data would have been cleaner and may have revealed stronger findings if collected in a lab setting, these are challenges facing schools, and the fact that these children still saw learning benefits despite these challenges is promising.

Finally, the study findings are important for math educators. These computer sessions were short, allowed children to practice estimation in a meaningful way not typically addressed in the classroom, had promising effects on children's learning, and were inherently engaging for the children, thus making the activity a great option to supplement the teacher's current math curriculum. Teachers should actively seek educational technology that is designed based on cognitive theories and encourage deep thinking rather than drill.

Videos from the computer sessions show every single child's face light up in a smile when their answer is right on or nearly right on the target number. Similarly, when children just barely find the target number within their range, it was not uncommon for the child to tap the researcher to show her how close he was or let out a sigh of relief. Structuring it as an estimation task also impacted children's perceptions of the activity. While walking one girl back to her classroom after her final computer session, the researcher asked, "Do you ever play any other math games on the computer?" to which the child laughed and replied, "Well, this isn't really a math game." Curious, the researcher asked why not and the child responded with, "In math you can only get the right answer, but in this game there isn't really a right answer. Well, there *is* a right answer but you don't have to get it exactly." Although this is anecdotal, it illustrates how the estimation task can engage children to explore important ideas about numbers and their magnitudes outside of the realm of "formal" mathematical concepts and algorithms.

Next Steps

In addition to avenues for future research discussed above, findings from this study can be used to improve the design of MANL. Firstly, using the data from this study, it may be possible to use Educational Data Mining or Learning Analytics methods to program MANL to detect children who are using the UDR tool ineffectively, and who are thus at risk for not seeing learning benefits from the activity (Baker & Yacef, 2009). After successful detection, the computer can take away the UDR feature and require the child to use a FR until reaching a level of mastery with the task before then moving back to UDR or provide additional instructions or scaffolding to help the child use the activity effectively. Identifying these children will be important for teachers as well, since this study found evidence to suggest that the children who are most likely to use the tool ineffectively have low number sense.

Another important next step to advance the effectiveness of MANL will be to explore how to harness the learning benefits to aid children in formal math abilities. Teachers can do this simply by encouraging approximation in classroom math activities in addition to formal algorithms. Because this study, as well as others, showed clear benefits to explaining one's thinking in mathematical problem solving, teachers could have groups of students discuss how they used approximation to solve specific problems. Using our earlier example, there are many methods to approximate quickly determining a tip (moving the decimal over one place and doubling it, doubling the sales tax and adding a bit more, or tipping one dollar for every seven of the bill) and dividing a restaurant bill among friends. Although the adults at the table may not appreciate the different problem solving strategies or see serious learning games from the practice, it is easy to imagine how a similar scenario could work in a formal classroom setting.

Limitations

Despite attempts to carefully design the study, there are certain limitations to address. Firstly, as discussed, improvements to the measures by including a non-symbolic quantity discrimination measure, adding more items and capturing response time in the Mental Number Line Task, collecting observed as well as expressed strategy on the Number Line estimation posttest, would have helped better explain the relationship between MANL and number sense. Secondly, although I wanted the research sessions to be relatively short, many of the children's second MANL session was less than ten minutes. Therefore, I could have required children to complete more trials, and ideally more sessions, which would have shed more light on the effect of target number on Trial PAE and potentially deepened the impact of MANL on number sense.

It could be argued that Trial PAE is not the best outcome variable to use as a measure of success on MANL, particularly for the MA UDR condition. Children in this condition could be focusing less on the placement of the line in the middle of their range (despite the fact that children are told to try to place as close to the target number as they can) because they know as long as the target falls anywhere within their range, they will be considered correct. Although this kind of approach is possible, I correlated the child's first click on the number line with the target number and found a strong positive correlation, indicating that children were at least partly attempting to place the middle line close to the target number. Luckily, the detailed log files will allow further examination of this concern.

Finally, one could argue that the Reading comparison group was not an appropriate control and I should have opted instead for another math software. The goal of this study was to explore the relationship between number sense and MANL, while utilizing the reading comparison group to control for the experience of working one-on-one with a researcher. Although our data cannot say conclusively how MANL stacks up against other existing math interventions, it showed that MANL provides promising learning opportunities to second-graders from schools serving low-income populations that are typically at risk for low math achievement.

Footnotes

[1]When running the same model without the Number Sense pretest factor, a significant interaction between Condition and whether the school was located in Chinatown or Harlem remained significant in the model, $F(2,51) = 3.41$, $p = .041$. Pairwise comparisons using Tukey's HSD reveal that in Harlem schools, children in the MA UDR condition are doing significantly better than children in the reading comparison group (adjusted $p = 0.039$), whereas condition has no effect within the Chinatown school.)

[2]I also ran all three analyses as binary logistic regressions using the appropriate outcome (Counting at least once at posttest, Number Relations at least once at posttest, Explicit explanation at posttest as the outcome) with condition as a factor, controlling for school location and gender, with stepwise removal of non-significant interactions. Even controlling for other factors, condition remained significant in the counting and explicit analyses. For Number Relations, although the main effect of condition is N.S., pairwise comparisons using the Bonferroni procedure show that children in the MA UDR condition are more likely to utilize the Number Relations strategies than the reading comparison group (adjusted $p = .017$).

[3]This was used as a measure of response time rather than the categorical researcher codes of quick or not, as regression is better suited for continuous variables. While the categorical variable of whether the trial was coded as quick or not significantly predicts Trial PAE in the regression model, using the continuous variable has several benefits. The beta coefficient gives more evidence to how response time impacts Trial PAE, and as it comes directly from the log file rather than researcher codes, it is a more reliable measure.

References

Anderson, D. R., & Kirkorian, H. L. (in preparation). Cognitive development through media.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*.

Baroody, A. J. (1990). How and when should place-value concepts be taught? *Journal for Research in Mathematics Education*, *21*(4), 281- 286.

Baroody, A. J., M. Lai, Mix, K. (2006). The development of young children's early number and operation sense and its implications for early childhood education. Handbook of research on the education of young children. B. Spodek and O. Saracho. Mahwah, NJ, Erlbaum. 2: 187-221.

Barth, H.C., & Paladino, A.M., (2011). The development of numerical estimation: evidence against a representational shift. Developmental Science, 14(1), 125-135.

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

Blair, C., & Razza, R. P. (2007). Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten. Child Development, 78(2), 647-663.

Booth, J. L. & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. Developmental Psychology, 41(6), 189-201.

Brankaer, C., Ghesquière, P., & De Smedt, B. (2014). Children's Mapping between Non-Symbolic and Symbolic Numerical Magnitudes and Its Association with Timed and Untimed Tests of Mathematics Achievement. *PloS one*, *9*(4), e93565.

Bull & Scerif (2001). Executive functioning as a predictor of children's mathematics ability: inhibition, switching, and working memory. Developmental Neuropsychology 19(3): 273-293.

Butterworth, B. & Reeve, R. (2008), Verbal counting and spatial strategies in numerical tasks. Evidence from indigenous Australia. Philosophical Psychology, 21(4), 443-457.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological bulletin*, *125*(3), 367.

Clearleft Limited (2010). Silverback (Version 2.0) [Computer application software]. Available from http://silverbackapp.com/

Clements, D. H. and J. Sarama (2007). Early childhood mathematics learning. Second Handbook of Research on Mathematics Teaching and Learning. F. K. Lester. Charlotte, NC, Information Age Publishing: 461-555.

Clements, D.H., and Sarama, J. (2008). Experimental evaluation of the effects of a research based preschool mathematics curriculum. American Educational Research Journal, 45, 443-494.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and cognitive processes*, *27*(4), 594-610.

Creighan, S. R., (in preparation). A Guide to Formative Research: A Case Study from MathemAntics.

Cross, C. T., Woods, T. A., & Schweingruber, H. (Eds.). (2009). Mathematics Learning in Early Childhood: Paths Toward Excellence and Equity. Washington, DC: National Academy Press.

Dehaene, S. (1997). The number sense: How the mind creates mathematics. Oxford, England, Oxford University Press.

Dehaene, S., & Akhavein, R. (1995). Attention, automaticity, and levels of representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 314.

Dehaene, S. Piazza, M., Pinel, P., Cohen, L. (2003). Three parietal circuits for number Processing, Cognitive neuropsychology 20(3-6), 487-506.

Dowker, A. (2003). Young children's estimates for addition: The zone of partial knowledge and understanding. In A. J. Baroody & A. Dowker (Eds.), The development of arithmetic concepts and skills: Constructing adaptive expertise (pp. 243 – 265). Mahwah, NJ: Erlbaum.

Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A., Klebanov, P., Pagani, L.S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., & Duckworth, K. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428-1446.

Dweck, C.S., & Leggett, E.L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256-273.

Dyson, N. I., N. C. Jordan, et al. (2011). A Number Sense Intervention for Low-Income Kindergartners at Risk for Mathematics Difficulties. Journal of Learning Disabilities.

Feigenson, L., Dehaene, S., Spelke, E., (2004). Core systems of number, Trends in Cognitive Sciences, 8, (7), 307-314.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & cognition*, *38*(7), 951-961.

Fennema, E., Carpenter, T.P., Jacobs, V.R., Franke, M.L., & Levi, L.W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. Educational Researchers, 27(5), 6-11.

Flagg, B.N. (1990). Formative Evaluation in Decision Making. In B.N Flag (Eds.), Formative research for educational technology (pp. 13 – 32). Hillside, NJ: Lawrence Erlbaum Associates, Inc.

Gallistel, C. R. (1990). The Organization of Learning. Cambridge, MA: Bradford Books/MIT Press.

Gallistel, C. R., and Gelman, R. (1992). Preverbal and verbal counting and computation. Cognition, 44, 43-74.

Gallistel, C. R., & Gelman, R. (2005). Mathematical Cognition.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature, 447*(7144), 589–591.

Griffin, S. (2004). Number Worlds: A research-based mathematical program for young children. In D. H. Clements & A.-M. DiBiase (Eds.), Engaging young children in mathematics: Standards for early childhood mathematics education (pp. 325-342). Mahweh, NJ: Lawrence Erlbaum Associates.

Gee, J.P. (2005). Learning by Design: good video games as learning machines. E-Learning, 2(1), 5-16.

Gelman, R., & Gallistel, C. R. (1986). The child’s understanding of number. Cambridge, MA: Harvard University Press.

Ginsburg, H. P. (1989). Children's arithmetic (2nd ed.). Austin, TX: Pro-Ed.

Ginsburg, H. P. (2006). Mathematical play and playful mathematics: A guide for early education. In D. Singer, R. Michnick-Golinkkoff & K. Hirsh-Pasek (Eds.), *Play = learning: How play motivates and enhances children's cognitive and social-emotional growth (Vol. 1).* Oxford, England: Oxford University Press.

Ginsburg, H. P. and A. J. Baroody (1990). The test of early mathematics ability: Second Edition. Austin, TX, Pro Ed.

Ginsburg, H. P., Pappas, S., Lee, Y-S, & Chiong, C. (2011). How did you get that answer? Computer assessments of young children's mathematical minds. In P. Noyce (Ed.). New Frontiers in Formative Assessment. Cambridge MA: Harvard University Press.

Ginsburg, H. P., Carpenter, K. K., & Labrecque, R. (2011). Introduction to MathemAntics: Software for children from age 3 to grade 3. In *International Society for Design and Development Conference*. Boston, MA.

Ginsburg, H.P., Carpenter, K., Labrecque, R., & Pagar, D. (in press). New Possibilities for Early Mathematics Education: Cognitive Guidelines for Designing High-Quality Software to Promote Young Children's Meaningful Mathematics Learning Anne Dowker (Ed). Oxford Handbook of Numerical Cognition.

Ginsburg, H., Jamalian, A., Creighan, S. (2013). Cognitive guidelines for the design and evaluation of early mathematics software: The example of MathemAntics. Lyn English & Joanne Mulligan (Eds.) Reconceptualising early mathematics learning.

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.

Hyde, D.C., Khanum, S. & Spelke, E.S., (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92-107.

Jordan, N.C., Kaplan, D., Oláh, L.N., and 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.

Laski, E. V. & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. Child Development, 78(6), 1723-1743.

Le Corre, M., Van de Walle, G., Brannon, E.M., and Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of counting as a representation of the positive integers. Cognitive Psychology, 52(2), 130-169.

Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., & Gunderson, E. A. (2010). What counts in the development of young children's number knowledge?. Developmental psychology, 46(5), 1309.

Li, J. (2002). Learning models in different cultures. Learning in culture and context: Approaching the complexities of achievement motivation in student learning. J. Bempechat and J. G. Elliott. San Francisco, Jossey-Bass: 45-63.

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.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), Aptitude, learning, and instruction: Vol. 3. Cognitive and affective process analysis (pp. 223-253). Hillsdale, NJ: Lawrence Erlbaum Associates.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. Handbook of research on mathematics teaching and learning. D. A. Grouws. New York, Macmillan: 575-596.

McCrink , K. & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science,* 15(11), 776-781.

Mix, K.S., (2009). How Spencer made number: First uses of the number words. *Journal of Experimental Child Psychology*, 102, 427–444.

Mix, K.S., (2010). Spatial tools for mathematical thought. In K.S. Mix, L.B. Smith, & M. Gasser (Eds.) *The Spatial Foundations for Language and Cognition* (pp. 41-66). Oxford: Oxford University Press.

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). Quantitative development in infancy and early childhood. *Infant Child Development* (Vol. 12, pp. 110-112 ST – Quantitative development in infancy). Oxford University Press.

Moeller, K., Pixner, S., Kaufmann, L. & Nuerk, H.-C. (2008). Children's early mental number line: Logarithmic or decomposed linear? *Journal of Experimental Child Psychology*, 103(4), 503-515.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality.

Moyer, R. S., & Landauer, T. K. (1973). Determinants of reaction time for digit inequality judgements. *Bulliten of the Psychonomic Society*, 1(3), 167-168.

Moyer-Packenham, P.S., Salkind, G., & Bolyard, J.J. (2008). Virtual manipulatives used by K-8 teachers for mathematics instruction: Considering mathematical, cognitive, and pedagogical fidelity. *Contemporary Issues in Technology and Teacher Education*, 8(3), 202-218.

National Governers Association Center for Best Practices and Council of Chief State School Officers (2010). *Common Core State Standards*. Washington, DC.

Nuerk, H.-C. & Willmes, K., (2005). On the magnitude representations of two-digit numbers. *Psychology Science, 47*(1), 52-72

Opfer, J. E., Siegler, R. S., Young, C. J. (2011). The powers of noise-fitting: reply to Barth and Paladino. *Developmental Science*, 14, 1194-1204.

Parmar, P., Harkness, S., & Super, C. M. (2004). Asian and Euro-American parents' ethnotheories of play and learning: Effects on preschool children's home routines and school behaviour. *International Journal of Behavioral Development, 28*(2), 97-104.

Petitto, A.L., (1990). Development of numberline and measurement concepts. *Cognition and Instruction, 7*(1), 55-78.

Pintrich, P.R. (2003). A motivational science perspective on the role of the student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667-686.

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.

Ramani, G., Siegler, R., Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *Journal of Educational Psychology, 104*(3), 661-672.

Sarama, J. & Clements, D. H., (2009). "Concrete" computer manipulatives in mathematics education. *Child Development Perspectives, 3*(3), 145-150.

Saxe, G.B., Earnest, D., Sitabkhan, Y., Halder, L.C., Lewis, K.E., and Zheng, Y. (2010). Supporting generative thinking about the integer number line in elementary mathematics. *Cognition and Instruction, 28*(4); 433-474.

Siegler, R.S., & Booth, J.L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428-444.

Starkey, P., E. S. Spelke, et al. (1990). Numerical abstraction by human infants. *Cognition,* 36: 97-128.

Sun, W., & Zhang, J. (2010). Teaching addition and subtraction facts: A Chinese perspective. *Teaching Children Mathematics, 8*(1), 28-31.

The Learning Company (2006). Reader Rabbit: Learn to Read with Phonics, Grade 1-2. Available from amazon.com.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499-535.

Vygotsky, L. S. (1986). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA:

The MIT Press.

Warschauer, M., Knobel, M., & Stone, L. (2004). Technology and equity in schooling: Deconstructing the digital divide. *Educational Policy*, *18*(4), 562-588.

Warschauer, M. & Matuchniak, T. (2010) New technology and digital worlds: Analyzing evidence of equity in access, use, and outcomes. *Review of Research in Education* (34).179-225.

Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S. K., Cohen, L., & Cohen, D. (2006). Principles underlying the design of" The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions, 2*(1), 19.

Wynn, K. (1995). Infants possess a system of numerical knowledge. Current *Directions in Psychological Science 4*: 172-177.

Xu, F. and E. Spelke (2000). Large number discrimination in 6-month-old infants. *Cognition 74*: B1-11.