INVESTIGATING EARLY SPATIAL AND NUMERICAL SKILLS IN JUNIOR KINDERGARTEN CHILDREN LEARNING IN AN INQUIRY- AND PLAY-BASED ENVIRONMENT

by

Ashley L. S. Olver

A thesis submitted in conformity with the requirements for the degree of Masters of Arts
Graduate Department of Applied Psychology and Human Development
Ontario Institute for Studies in Education
University of Toronto

© Copyright by Ashley Olver, 2013
INVESTIGATING EARLY SPATIAL AND NUMERICAL SKILLS IN JUNIOR KINDERGARTEN CHILDREN LEARNING IN AN INQUIRY- AND PLAY-BASED ENVIRONMENT

Masters of Arts, 2013
Ashley L. S. Olver

Department of Applied Psychology and Human Development
Ontario Institute for Studies in Education
University of Toronto

Abstract

In the current study, three possible interpretations of children’s number line estimation (NLE) performance were examined for appropriateness and possible correlates of performance were tracked over time in a classroom exemplifying recommended mathematics pedagogy for young children. In December and May, 21 4-year-olds completed the NLE task (0-10 range) and measures of numerical knowledge, spatial skills, and visual-motor integration. With high-quality teaching, children made large gains in these skills ($d = 0.96–1.28$). Due to uniformly high achievement, few expected correlations were observed, however. A strategy account of NLE performance was supported over the traditional logarithmic-to-linear shift account and the newly proposed proportion-judgement account. Patterns of error in estimation provide a better indication of understanding of the linear number line than models of best fit. Indeed, interpreting linearity of NLE as indicative of an underlying representation of number could lead to inappropriate conceptualizations of math learning disabilities and misguided interventions.
Acknowledgements

When I began my Masters Degree, a thesis felt like an overwhelming and insurmountable challenge. There are so many people that I want to thank for helping me overcome that feeling and create something I am proud of.

First and foremost, I am thankful for my extraordinary husband, Derek, who loves me when I am hard to love; believes in me when I don’t believe in myself; and carries more than his fair share when my load becomes too heavy. I don’t know a better man. I am grateful for my tiny loves, Joey and Karl, and their companionship during long days of writing. I am thankful to all of my friends and family, but especially my dear friend Julie Bell, for sticking with me through the highs and lows of graduate school and waiting for me on the other side. I am also grateful to have had the opportunity to learn alongside and from some of the most talented and committed people I know over these past two years: Kristina, Sharon, Kathleen, Alan, Linda, and Amanda. Thanks for the memories!

I would like to thank my supervisor, Dr. Joan Moss, who has boundless enthusiasm for understanding how children think about and learn mathematics, for providing me with the resources and support I needed to complete this project. I would also like to thank my second reader, Dr. Yukari Okamoto, for her thoughtful feedback throughout the project. Many, many thanks go to Zachary Hawes, who helped with everything from data collection to reading drafts and who was a wonderful sounding board. His passion for research is contagious!

I am deeply appreciative of the opportunity to conduct my research at the laboratory school. Thank you to the classroom teacher for welcoming me into your classroom; allowing me the opportunity to observe exceptional teaching and learning in action; and spending time discussing this research with me. The children were a joy to work with; I must express gratitude to them and to their parents for allowing them to participate.

Last but not least, I need to thank Dr. Brenda Smith-Chant and Dr. Deborah Kennett for encouraging me to pursue graduate studies and for providing me with the research opportunities and recommendations that got this journey started.
# Table of Contents

Abstract .................................................................................................................................................. ii
Acknowledgements ................................................................................................................................iii
Table of Contents ................................................................................................................................... iv
List of Tables .......................................................................................................................................... vi
List of Figures .......................................................................................................................................... vii

## Introduction

- Early Numerical Skills .................................................................................................................. 2
- Early Spatial Skills .......................................................................................................................... 12
- Visual-Motor Integration and Number Line Estimation ................................................................. 17
- Malleability of Early Numerical and Spatial Skills ........................................................................ 19
- Current Study ..................................................................................................................................... 22

## Methods

- Participants .......................................................................................................................................... 25
- Classroom ............................................................................................................................................. 26
  - Sustained Play: Example 1 ............................................................................................................... 28
  - Sustained Play: Example 2 ............................................................................................................... 29
  - Teacher-Directed Instruction: Example 1 ...................................................................................... 30
  - Teacher-Directed Instruction: Example 2 ...................................................................................... 32
- Materials ............................................................................................................................................. 33
  - Number Line Estimation Task ....................................................................................................... 33
  - Number Knowledge Test (Abridged) ............................................................................................. 35
  - Children’s Mental Transformation Test ......................................................................................... 37
  - Beery VMI (Short Form) ................................................................................................................ 37
- Procedure ........................................................................................................................................... 38
- Data Analysis ....................................................................................................................................... 39
  - Model Selection .............................................................................................................................. 40
  - Linearity of Estimation Patterns ................................................................................................... 41
  - Strategy Use ...................................................................................................................................... 42
  - Relationships amongst Variables ................................................................................................. 42
The Sample and Development...........................................................................43

Results..........................................................................................................................43

- Model Selection........................................................................................................44
- Linearity of Estimation.................................................................................................44
- Strategy Use................................................................................................................45
- Relationships amongst Variables...............................................................................46
- The Sample and Development....................................................................................48

Discussion.....................................................................................................................50

- Does a Strategy Account Better Explain NLE Performance than a Logarithmic-to-Linear Shift Account or a Proportion-Judgment Account? .................................................................50
- How are Numerical Knowledge, Spatial Skills, and NLE Performance Related? ..........54
- Is VMI a Confound in the NLE task or a Correlate that Increases its Value in Predicting other Math Abilities? ..............................................................................................................57
- Description of the Development of Children Learning in a Play- and Inquiry-Based Classroom....................................................................................................................................58
- Future Research and Conclusions..............................................................................59

References.....................................................................................................................62
List of Tables

Table 1. Math Topics Intentionally Addressed between December and May in the Areas of Numerical and Spatial Skills...........................................................................................................................................72

Table 2. Descriptive Statistics for All Variables................................................................................73

Table 3. Summary of Number Line Estimation Data for Individual Children at Time 1........74

Table 4. Summary of Number Line Estimation Data for Individual Children at Time 2........75

Table 5. β Values for NLE Patterns of Children that were Best Fit by an Unbounded Power Function at each Time Point and the Direction of the Change.......................................................76

Table 6. Correlations between Variables at Time 1 and Time 2......................................................77

Table 7. Comparisons of Strategy Groups on Spatial, Numerical, and VMI Skills at Time 1 and Time 2....................................................................................................................78
List of Figures

Figure 1. Five potential models of best fit applied to the number line estimation task........79

Figure 2. A sample of the variety of play materials available in the classroom..................80

Figure 3. Examples of children’s use of play materials....................................................81

Figure 4. Examples of creations built cooperatively during sustained free play.................82

Figure 5. Graphing activity.................................................................................................83

Figure 6. Materials used in the number line estimation task..............................................84

Figure 7. The lines of best fit for number line estimation patterns of two hypothetical children with the same $R^2_{lin}$.................................................................85

Figure 8. Number line estimation patterns for individual children....................................86

Figure 9. Group level number line estimation data at Time 1 (T1) and Time 2 (T2).........88

Figure 10. Mean PAE across target number for groups of individuals with Number Line Estimation patterns categorized in each of the five strategy categories at each time point.................................................................89

Figure 11. Scatter plots of the relationships between numerical skills (NKT-A) and VMI (Beery VMI) within and between times points....................................................90

Figure 12. Varying qualities of hash marks made on the number line estimation task........91
Investigating Early Spatial and Numerical Skills in Junior Kindergarten Children Learning in an Inquiry- and Play-Based Environment

A great deal of evidence across disciplines has supported experiences in the early years as critically important to lifelong development in the areas of learning, behaviour, and mental and physical health (McCain, Mustard, & McQuaig, 2011). In particular, it has recently been highlighted how vital early mathematics learning is to later academic achievement in both mathematics and reading (Duncan et al., 2007). In a metanalysis of six large, longitudinal data sets, Duncan et al. (2007) uncovered that numeracy skills at school entry were a better predictor of later academic achievement than early reading skills and attention, when overall cognitive ability, behaviour, and other background characteristics were controlled for. In line with this finding, the National Council of Teachers of Mathematics and National Association for the Education of Young Children (NCTM & NAEYC, 2010) acknowledged the importance of a high-quality, challenging math education for young children. Specifically, the NCTM (2000) proposed that the core areas of math teaching in Kindergarten to Grade 2 should be numeracy (e.g., counting, understanding magnitude of numbers, composing and decomposing number) and geometry/spatial skills (e.g., investigating the composition and decomposition of two- and three-dimensional shapes; recognizing symmetry; conducting slides, flips, and rotations; visualizing shapes and spatial relations). Unfortunately, math teaching is often minimized and overshadowed by a focus on social development and literacy in the education of young children (Lee & Ginsburg, 2007a; Phillips, Gormley, & Lowenstein, 2009). When math is addressed, spatial skills are often neglected (Clements & Sarama, 2011). Seemingly, some educators may not be aware of the breadth and depth of mathematical thinking that young children are ready for (Lee & Ginsburg, 2007b). There are, undoubtedly, exceptional classrooms where math is taught in a fashion consistent with the best practices recommended by national organizations (NCTM, 2000; NCTM & NAEYC, 2010). It is valuable for educators and researchers alike to document examples of such classrooms. In the current study, such a classroom was described and the development of the early spatial and numerical abilities of the children who learn there were investigated.

Following is a review of literature in the areas of early numerical and spatial skills that led to the design of the current study. Three themes permeate the review. One theme centers on how numerical and spatial skills relate to one another and how each relates to other math
abilities more broadly. A second theme centers on the methodological concerns and
definitional inconsistencies in both early numerical and spatial skills research. The last theme
centers on the malleability of these skills. These themes will be explored using the Number
Line Estimation task (Siegler & Opfer, 2003), an assessment task that has received an
abundance of attention in the mathematical cognition (e.g., Barth & Paladino, 2011; 
Dehaene, Izard, Spelke, & Pica, 2008; Laski & Siegler, 2007; Siegler & Booth, 2004) and
educational intervention research (e.g., Siegler & Ramani, 2008; Whyte & Bull, 2008) over
the past decade, as a focal point.

Early Numerical Skills

A growing number of studies have shown that the numeracy skills that young children
begin formal schooling with are strongly predictive of their later math achievement (Aunola,
Leskinen, Lerkkanen, & Nurmi, 2004; Duncan et al., 2007; Jordan, Glutting, & Ramineni,
2010; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Krajewski, & Schneider, 2009; 
Passolunghi & Lanfranchi, 2012). For example, Jordan and colleagues found that basic
numeracy skills measured in Kindergarten were predictive of math achievement and growth
in math ability in Grades 1 through 3 (Jordan et al., 2009) and that the influence of these
basic numeracy skills remained after controlling for age and cognitive abilities (i.e.,
vocabulary, non-verbal reasoning, and memory; Jordan et al., 2010). Additionally,
Mazzocco & Thompson (2005) found that numeracy skills measured in kindergarten (i.e.,
reading numerals, understanding number constancy, making magnitude judgments of one-
digit numbers, mental addition of one-digit numbers) were predictive of a math learning
disability in Grade 3. Thus, research supports that early numeracy skills are predictive of
both math achievement and math difficulty.

Despite a clear indication that early numeracy skills are important, there is no agreed
upon definition of early numeracy or “number sense” between the disciplines of psychology
and education or by researchers within these disciplines (Berch, 2005; Gersten, Jordan, & 
Flojo, 2005). In a review of the literature, Berch (2005) compiled a list of 30 components
that have been proposed to make up the construct of number sense, ranging from an intuition
or feeling for number to understanding the base-ten number system. Until recently, there
appeared to be mounting agreement that one key early numeracy skill is the development of
an internal linear number line representation for a small numerical range (e.g., Case & Okamoto, 1996; Jordan et al., 2009; Krajewski & Schneider, 2009; Laski & Siegler, 2007). Case & Okamoto (1996) proposed that the linear number line representation is the central conceptual structure for understanding number. They described it as a mental construct where knowledge of number symbols and number words are linked to the quantities they represent, and these quantities are arranged in ascending order, equally spaced, from left-to-right (Case & Okamoto, 1996).

The most common way to determine the linearity of children’s number line representations is a bounded number line estimation (NLE) task (e.g., Siegler & Opfer, 2003). In this task, children are presented a line with 0 (or 1; see e.g., Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Whyte & Bull, 2008) anchoring the left end and a larger number (e.g., 10, 100) anchoring the right end. Children are given a number between these points and asked to mark where it belongs on empty number line. A fresh line is presented for each number. The number representing the midpoint (e.g., 5, 50) is sometimes used as a teaching trial (e.g., Barth & Paladino, 2011; Booth & Siegler, 2006; Siegler & Booth, 2004). The estimates of children ($y$) are plotted against the actual numbers presented to them (target number; $x$) and these scatters are analyzed to determine whether a linear function is a good fit for the patterns of performance seen (see Figure 1a). The goodness-of-fit statistic typically used is the coefficient of determination ($R^2$), which indicates the proportion of variance in responses that can be accounted for by the model (e.g., linear). If a linear function is a good fit, children are typically assumed to have developed an internal linear representation for the numbers within the tested range (e.g., Berteletti et al., 2010; Siegler, Thompson, & Opfer, 2009).

Like early numeracy skills in general, linearity of NLE performance, specifically, has been found to be positively correlated with overall math ability of young children (Booth & Siegler, 2006; Siegler & Booth, 2004). Linearity of NLE performance has also been found to be positively related to other specific aspects of math ability, such as accuracy of magnitude comparison (Siegler & Laski, 2007); computational, measurement, and numerosity estimation (Booth & Siegler, 2006); existing addition knowledge (Booth & Siegler, 2008; Siegler & Mu, 2008); the ability to learn new addition combinations (Booth & Siegler, 2008);
and understanding of the principles of counting (Fischer, Moeller, Bientzle, Cress, & Nuerk, 2011). Further, children with math difficulties or math learning disabilities have been shown to produce patterns of performance on NLE tasks that are less linear than their same age peers (Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007; Geary, Hoard, Nugent, & Byrd-Craven., 2008).

These findings have not been disputed. What has been disputed is whether linear performance on an NLE task is indicative of an internal linear representation for numbers and how to best analyze and interpret non-linear overall patterns of performance (e.g., see Barth & Paladino, 2011 and subsequent replies from Opfer, Siegler, & Young, 2011 and Barth, Slusser, Cohen, & Paladino, 2011; Boumeester & Verkoeijen, 2012). In a large body of work by Siegler and colleagues (e.g., Booth & Siegler, 2006; Laski & Siegler, 2007; Siegler & Opfer, 2003; Siegler & Ramani, 2008) and others (e.g., Berteletti et al., 2010; Fischer, 2003; Halberda & Feigenson, 2008) children’s patterns of performance on NLE tasks were interpreted as corresponding to their internal representations of numbers; patterns of performance were analysed for linear and logarithmic fits (see Figure 1b), and thereby, linear and logarithmic representations of number.

The motivation for fitting data to a logarithmic model opposed to other models (e.g., exponential, quadratic) comes from a generally accepted theory that humans posses an innate system for approximating the magnitude of numbers that is logarithmically compressed (Feigenson, Dehaene, & Spelke, 2004; Dehaene et al., 2008; but see Núñez, 2011). This system has often been described as a “mental number line” (e.g., Restle, 1970; Dehaene, Bossini, & Giraux, 1993) where relatively smaller numbers are typically represented on the left and relatively larger numbers on the right (Dehaene et al., 1993; Fischer, 2008). Unlike with a linear representation of number, though, magnitudes are not equally spaced on this innate mental number line; magnitudes are represented logarithmically (according to the Weber-Fechner law) such that the likelihood of confusing a number with an adjacent number increases as the magnitude of the number increases (Feigenson et al., 2004). As such, the distance between numbers of smaller magnitudes is exaggerated (i.e., easier to discriminate) and the distance between larger magnitudes is minimized (i.e., more difficult to discriminate). It has been proposed that this logarithmic mental number line is used,
automatically, across the lifespan (Halberda & Feigenson, 2008) and that over time this logarithmic representation is complimented, but not replaced, by linear representations for familiar number ranges (e.g., Berteletti et al., 2010; Siegler et al., 2009). This view has been referred to as an “overlapping waves” theory (see Chen & Siegler, 2000).

A great deal of data has been interpreted in support of a logarithmic-to-linear shift in underlying mental representations of magnitude (e.g., see Figures 1 and 2 in Opfer et al., 2011). For example, using a 0-100 number line, Siegler and Booth (2004) found that a logarithmic model was a better fit than a linear model for the median estimates of Kindergarten children ($R_{lin}^2 = .49$ vs. $R_{log}^2 = .75$) but that a linear model was a better fit estimates of Grade 2 students ($R_{lin}^2 = .95$ vs. $R_{log}^2 = .88$; see Siegler & Booth, 2004, Figure 2). Behaviourally, this means that, on average, Kindergarten children in this study tended to overestimate the magnitude of small numbers and underestimate the magnitude of larger numbers. Siegler and Booth (2004) interpreted this to mean that these young children were drawing on a logarithmic mental number line, where smaller magnitudes are easier to discriminate than larger magnitudes, and therefore are given more weight, whereas Grade 2 students were drawing on equally spaced linear representations.

Across studies, whether a linear model is the best fit for patterns of responses on NLE tasks has been shown to be partially dependent on age and on the range of numbers for which children are assessed. In general, linear fits for small number ranges (i.e., 1-10, 1-20) have been seen first, between 3 and 6 years of age (Berteletti et al., 2010; Whyte & Bull, 2008) and linear fits for larger ranges have been seen later in development, with the 0-100 range showing a linear fit between 6 and 8 years of age (Siegler & Booth, 2004), and the 0-1000 range showing a linear fit between 8 and 12 years of age (Siegler & Opfer, 2003). It should be noted that the pattern of shifts in models of best fit described here are based on group averages and substantial individual differences in linearity of NLE responses have been observed amongst same-aged children across studies (e.g., Berteletti et al., 2010; Geary et al., 2008; Laski & Siegler, 2007). What accounts for these individual differences is important to consider and will be discussed further later.
While this logarithmic-to-linear shift account is popular, alternative interpretations of the data have been made. One alternative interpretation is a dual-linear to single-linear shift in representation (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008; Moeller, Pixner, Kaufmann, & Nuerk, 2009). In this account, the curve seen in logarithmic functions applied to patterns of NLE is seen as better conceptualized as the point at which two linear functions, a linear function with a steep slope and a linear function with a shallow slope, meet.

Ebersbach et al. (2008) proposed that the pattern of children’s NLE aligns with a steep linear function for the range of numbers they are familiar with and a shallower linear function for the less familiar range. In support of this theory, Ebersbach et al. (2008) found that a dual-linear model was a better fit for the NLE (0-100 and 0-1000 ranges) of 5- to 9-year-old children than a logarithmic or single-linear model, and that the point at which the two linear functions met (i.e., breakpoint) was correlated with the number range in which a child began to experience difficulty in a counting task. Similarly, Moeller et al. (2009) suggested that children’s representation of two-digit numbers is less accurate than their representation for one-digit numbers, and therefore two-digit numbers produce a shallower sloping linear function than one-digit numbers. Unlike Ebersbach et al. (2008), Moeller et al. (2009) predicted that the breakpoint will always occur at 10. They suggested that the whole range becomes more linear when children’s understanding of tens and ones are integrated. Moeller et al. (2009) found that both the logarithmic and dual-linear model fit the NLE (0-100 range) of 5- to 7-year-old children well, but in support of their theory, they also found that the modal breakpoint was around 10 (i.e., between 6 and 15). What logarithmic-to-linear shift and dual-linear to single-linear shift accounts of NLE patterns have in common is an assumption that patterns of performance on this task are directly indicative of underlying internal representations of numerical magnitude.

Contrary to this position, Barth and colleagues (e.g., Barth & Paladino, 2011; Barth, Slusser, Cohen, & Paladino, 2011; Slusser, Santiago, & Barth, 2013) and others (e.g., Boumeester & Verkoeijen, 2012; Cohen & Blanc-Goldhammer, 2011) have argued that NLE patterns cannot be interpreted as being the same as underlying internal representations of numerical magnitude. In fact, Barth and Paladino (2011) proposed that the NLE task does not assess a child’s understanding of a single numerical magnitude at all. Rather, they characterized the NLE task as a proportion judgment task, where the magnitude of a smaller
number (i.e., target number) must be estimated relative to the magnitude of a larger number (i.e., the upper anchor of the number line). Thus, in order to demonstrate an accurate, linear pattern of estimation on this task a child must have at minimum three kinds of numerical understanding; they must understand: i) the magnitude of the target number (\(N\)); ii) the magnitude of the upper anchor; and iii) how these magnitudes are related (Barth & Paladino, 2011). Additionally, the pattern of estimation produced is thought to be influenced by non-numerical competencies, namely spatial skills (Slusser et al., 2013). A child who understands the numerical proportional relationship must demonstrate this knowledge by marking the number line in the location that generates an equivalent spatial proportional relationship (Slusser et al., 2013). Therefore, the NLE task could be interpreted as a proportion judgement task at two different levels.

In line with their proportion-judgement account, Barth and colleagues predict that children’s patterns of performance on NLE tasks are best fit by the same functions that fit perceptual proportion judgement data (e.g., estimating the volume of a cylinder; estimating length of time intervals; Barth & Paladino, 2011); that is, they predict that NLE patterns are best fit by the cyclical power model (see Hollands & Dyre, 2000). The cyclical power model evolved from Steven’s power law. Steven’s power law describes an estimated stimulus magnitude (\(y\); here, the estimated position of \(N\) on the number line) as a function of an actual stimulus magnitude (\(x\); here, the actual position of \(N\) on the number line) using the equation

\[ y = \alpha x^\beta, \]

where \(\alpha\) is a scaling parameter and \(\beta\) represents the bias in estimating the magnitudes of particular stimuli (takes on different values for viscosity, brightness and so on). Steven’s power law has been extended to proportion judgements (Hollands & Dyre, 2000; Spence, 1990) by applying it to the estimation of both the part (here, \(N\)) and the whole (here, the upper anchor, \(A\)). What results is a equation where the scaling parameter is cancelled out leaving just one free parameter, \(\beta\):

\[ y = \frac{x^\beta}{x^\beta + (A - x)^\beta}. \]

This equation produces an S-shaped or reverse S-shaped pattern of over- and under-estimation. When \(\beta > 1\), values below the midpoint (0.5\(A\)) are underestimated and values between the midpoint and \(A\) are overestimated; conversely, when \(\beta < 1\), values below the
midpoint are overestimated and values above the midpoint are underestimated. Hollands and Dyre (2000) referred to this as a “one-cycle power model” (see Figure 1d) and they predicted it to result when the upper anchor point is used to make a proportion judgement. When additional imagined reference points are used, more than one S-shaped pattern is expected to result. For example, when the midpoint is used as an additional reference point, a “two-cycle power model” (see Figure 1e) is predicted. Oversampling target numbers from below the midpoint, as is often done (e.g., Ashcraft & Moore, 2012; Booth & Siegler, 2006; Siegler & Opfer, 2003), was noted to be an issue because it is done specifically to discriminate between logarithmic and linear patterns; as such, this practice allows overestimation below the midpoint to readily be seen but reduces the ability to see underestimation above the midpoint which is important in deciphering cyclical fits (Barth & Paladino, 2011; Slusser et al., 2013).

Barth and Paladino (2011) proposed that more accurate and linear-looking NLE patterns are a consequence of two key developments resulting from age and experience. One key development is a reduction in the bias in estimating the individual relevant magnitudes (e.g., target number, upper anchor), which is represented by $\beta$. The closer $\beta$ is to 1 the more linear the fit seen, such that when $\beta = 1$, a perfect linear fit is observed (i.e., $x = y$). A second key development is the use of additional reference points. Children who use both anchors as well as additional imagined reference points are expected to produce more linear-looking NLE patterns than those who use fewer reference points. In fact, they proposed that less accurate and more logarithmic-looking (or exponential-looking, when $\beta > 1$) NLE patterns may result from a poor understanding of the magnitude of the upper anchor and therefore an inability to use it as a consistent reference point. When the upper anchor is not used, the NLE task becomes an unbounded magnitude judgement and an “unbounded power function” (described by Steven’s power law; see Figure 1c) provides a good fit for the resulting estimation patterns. Slusser et al. (2013) pointed out that other factors could lead to inaccurate estimations, though, such as using inappropriate strategies (e.g., simply counting from left to right by randomly selected units that may be inconsistent from trial to trial) or a poor understanding of any of the target numbers. Given that there are multiple sources of the variability seen in responding, it does not make sense to interpret any pattern of responding as indicative of an underlying representation of number (Slusser et al., 2013).
The proportion-judgment account has been specifically testing in only a handful of studies (e.g., Barth & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011; Slusser et al., 2013). In these studies, the Akaike information criterion with a correction for small sample size (AICc) was often used as the goodness-of-fit statistic because it allows for comparison to be made across models with different numbers of parameters. Better fitting models have smaller AICc values. \( R^2 \) values were also reported because of general familiarity with the meaning of the statistic. Generally, in these studies, support was found for a proportion-judgement account over a logarithmic-to-linear shift account. For example, Slusser et al. (2013) presented the NLE task (0-20 to 0-100,000 ranges) to 5- to 10-year-old children. For the 0-100 range, models of best fit for average NLE patterns transitioned from an unbounded power function (5-year-olds) to a one-cycle power model (6-year-olds) to a two-cycle power model (7- and 8-year-olds). This suggested that older children were better able to make use of additional imagined reference points. Additionally, they found that \( \beta \) values for the models of best fit were correlated with age in months, suggesting that bias in estimating necessary magnitudes decreased with age.

Additionally, Slusser et al. (2013) pointed out that a great deal of past NLE data can be reinterpreted to be better fit by a proportion-judgement account (specifically, they direct readers to “see e.g., Booth & Siegler, 2006, Figure 2; Ebersbach et al., 2008, Figure 2; Laski & Siegler, 2007, Figures 1 and 2; Moeller et al., 2009, Figure 3; Siegler & Booth, 2004, Figure 3; Siegler & Mu, 2008, Figure 1”, p. 196). Indeed, the example from Siegler and Booth (2004) provided earlier could be viewed through a proportion-judgment lens. The patterns of over- and under-estimation seen in Kindergarten children and the more accurate estimation patterns seen in Grade 2 children were interpreted as indicating underlying logarithmic and linear representations of number, respectively, by Siegler and Booth (2004); however, a visual inspection of the data (Siegler & Booth, 2004, Figure 2) suggests a one-cycle power model with a \( \beta < 1 \) for Kindergarten children and a two-cycle power model with a \( \beta \) close to 1 for Grade 2 children may be good fits, which would be consistent with predictions generated from the proportion-judgement account (although these models would need to be compared statistically to determine the best fit).
Opfer, Siegler, and Young (2011) responded to the original work of Barth and Paladino (2011), defending the logarithmic-to-linear shift account and taking issue with a number of facets of the work of Barth and colleagues. The primary claim of Opfer et al. (2011) was that power functions are excessively flexible. They asserted that although the power model seems simplistic because it only has one free parameter ($\beta$), changes in the value of $\beta$ allow the function to fit essentially any data set. Barth, Slusser, Cohen, and Paladino (2011) counter that the cyclical power model is both “theoretically motivated” and “mathematically sensitive” unlike linear functions which are “insensitive to systematic fluctuations in the data” (p. 1205).

While Barth and colleagues have fairly adeptly responded to the criticism of Opfer et al. (2011; see Barth et al., 2011), two findings that are not predicted by the proportional-judgement account arose in one of their own studies (i.e., Slusser et al., 2013). First, on the 0-1000 and 0-100,000 ranges, average NLE patterns were best fit by a one-cycle power model for older children (10-year-olds) but a two-cycle model for younger children (8- and 9-year-olds). Second, the $\beta$ values for older children did not simply approach 1 (i.e., perfect linearity); rather, the $\beta$ values for older children eventually overshot 1, shifting from a pattern of over- and under-estimation to a pattern of under- and over-estimation. Slusser et al. (2013) acknowledged that similar patterns were observed by Cohen and Blanc-Goldhammer (2011) with adults and Ashcraft and Moore (2012) with children and adults, but they had difficulty explaining these unexpected findings.

Furthermore, although evidence of making use of anchor points and additional imagined reference points is plentiful (e.g., Ashcraft & Moore, 2012; Ebersbach et al., 2008; Siegler & Opfer, 2003; White & Szucs, 2012), the correspondence between one- and two-cycle power models and use of anchor points predicted from the proportion-judgment account has not been found. By analyzing mean percentage absolute error in estimating target numbers, Ashcraft & Moore (2012) and White and Szucs (2012) uncovered similar evidence for a progression in the use of anchor points. Using a 0-100 range, Ashcraft and Moore (2012) found that 6-year-old children showed low error and low variability at 0 with a steady increase in errors and variability through to 75. Error declined slightly approaching 100 but variability remained high. Strategically, it appeared that these children started
counting from the left-hand side, with some children making use of the endpoint. Less error overall was seen in the estimates of 7-year-old children compared to 6-year-old children. 7-year-old children showed low error and low variability at 0 and 100 with the greatest errors and greatest variability surrounding 50. Ashcraft and Moore (2012) described this as a tent-shaped pattern. It appeared these children had developed a more sophisticated strategy—starting from whichever end-point is closest. 8- to 10-year-old children and adults showed the lowest error overall. They showed low error and low variability at 0, 50, and 100, with error at 25 and 75 being slightly higher. Ashcraft and Moore (2012) described this as an M-shaped pattern. These participants seemed to be using the most sophisticated strategy—using both endpoints as well as an imagined midpoint to aid in estimation. White and Szucs (2012) found a similar pattern using a 0-20 range; on this more familiar range, 6-year-old children showed a tent-shaped error pattern while 7- and 8-year-old children showed an M-shaped error pattern. Both Ashcraft and Moore (2012) and White and Szucs (2012) found that M-shaped error patterns, denoting strategic use of an imagined midpoint were related to more linear NLE patterns. Importantly, however, while this M-shaped pattern was seen increasingly with age, a two-cycle model was not an increasingly better fit for NLE data with age. This is of concern because Barth and colleagues predicted that use of a midpoint in estimation should result in a two-cycle power model.

Both Ashcraft and Moore (2012) and White and Szucs (2012) found that younger children in their samples, who showed error patterns (e.g., tent-shaped) indicating less strategic sophistication, produced NLE patterns that could be fit by a variety of non-linear functions (e.g., logarithmic, one-cycle power, two-cycle power). Yet, their interpretations of these similar findings were quite different. Ashcraft and Moore (2012) cited the claims of Opfer et al. (2011) that power functions were excessively flexible to support their decision to ignore the fact that a one-cycle power model provided the best fit for the NLE data of 36% of 7-year-olds (linear = 32%; exponential$^1$ = 24%; two-cycle = 4%) and 40% of 8-year-olds

---

$^1$ Ashcraft and Moore (2012) used a number-to-position version of the NLE task, as opposed to the position-to-number version that has been described thus far. The tasks are identical except that children are provided with the mark on the number line and asked to provide the corresponding number, as opposed to being provided with the number and being asked to provide the mark. Ashcraft and Moore (2012) considered an exponential pattern on this task to be equivalent to a logarithmic pattern on the typical NLE task (as Siegler & Opfer, 2003 did). While not explicitly stated, this appears to be because an exponential function is the inverse of a logarithmic function and the direction of the task has been reversed.
(linear = 40%; exponential = 0%; two-cycle = 10%). Instead, they focused on the transition from an exponential function providing the best fit for most 6-year-olds (58%) to a linear function providing the best fit for 9- and 10-year-olds and adults (62%, 52%, and 35%, respectively; one-cycle fits remained high: 28%, 35%, and 55%, respectively). Mirroring Siegler and Opfer’s (2003) conclusions from nearly a decade earlier, Ashcraft and Moore (2012) proposed that the strategic patterns seen resulted from drawing on either logarithmic (tent-shaped patterns) or linear (M-shaped patterns) internal representations of number to complete the task. A counterpoint to this position is that linear and non-linear patterns result from strategy use and are not reflective of internal representations of number at all (as Barth and colleagues suggested). This was the conclusion of White and Szucs’ (2012), who further hinted that given the variability in strategy use, it may not be fruitful to discern the shape of non-linear fits at all.

Despite the flaws in the proportion judgment account that have been noted, what Barth and colleagues (e.g., Barth & Paladino, 2011; Barth et al., 2011; Slusser et al., 2013) and others (e.g., Boumeester & Verkoeijen, 2012; Cohen & Blanc-Goldhammer, 2011; White & Szucs, 2012) have brought to the forefront is that the NLE task may not be a “relatively pure measure of numerical estimation” as Siegler and Booth (2004, p. 428) claimed. Indeed, the spatial skills of children have been proposed to be relevant to the task by a number of researchers (e.g., Ebersbach et al., 2008; Slusser et al., 2013). Early spatial skills and their relevance to the NLE patterns of young children will be explored in the next section.

**Early Spatial Skills**

Spatial abilities have been found to be robustly linked to math achievement across domains of math (see Mix & Cheng, 2012 for a review). High math achievers in Grades 2 through 7 have been found to have greater spatial ability than low math achievers (Guay & McDaniel, 1977). Further, Jordan et al. (2010) found that the spatial reasoning of students in Grades 1 and 4 was related to their math achievement, after vocabulary, short term memory, working, and basic number sense were accounted for. More specifically, Cheng & Mix (2012) found that 6- to 8-year-olds who received mental rotation training demonstrated significantly improved performance on math tasks, and on missing term arithmetic problems in particular. Little research has been carried out to examine which specific spatial skills are
predictive of which specific components of mathematical skill (Mix & Cheng, 2012), as Cheng & Mix (2012) have.

A contributing factor to the paucity of this specific research is likely that, as with number sense (Berch, 2005), there is no agreed upon definition of spatial ability (Hegarty & Waller, 2005). While there is agreement that spatial ability is not a unitary construct, there is no consensus on the number or kinds of components that characterize it (Hegarty & Waller, 2005). For example, Linn & Petersen (1985) identified three components (i.e., spatial perception, mental rotation, and spatial visualization), whereas Carroll (1993) identified three major (i.e., spatial visualization, spatial relations, and flexibility of closure) and three minor (i.e., closure speed, perceptual speed, visual memory) spatial abilities. The research is further complicated by the fact that researchers refer to similar spatial abilities by different names and different spatial abilities by the same name (Hegarty & Waller, 2005). For example, mentally rotating a two- or three-dimensional figure quickly and accurately would be referred to as mental rotation by Linn & Petersen (1985) but as spatial relations by Carroll (1993). This makes it difficult to compare and build upon studies where different measures (e.g., those suitable for different age groups) and different theoretical models of spatial ability were used.

Uttal, Meadow, Tipton, Hand, Alden, Warren, & Newcombe (2013) recently proposed a classification system that unifies disparate definitions and categorizations of spatial abilities (e.g., those of Linn & Peterson, 1985 and Carroll, 1993). Uttal et al. (2013) classified spatial skills based on two dimensions, the kind of spatial information involved (intrinsic vs. extrinsic) and the kind of task being performed with that information (dynamic vs. static), resulting in four categories of abilities. Intrinsic information defines an object without regard for its surroundings, whereas extrinsic information refers to relationship amongst objects in a group. During a static task the spatial information remains stationary in the mind’s eye, without changing orientation, location, or dimension, whereas during an dynamic task the spatial information moves either physically or mentally. Thus, by Uttal et al.’s (2013) classification, mentally rotating a figure would be an intrinsic-dynamic task.

A limited amount of research has been conducted connecting early spatial abilities and early numeracy skills in young children (Mix & Cheng, 2012). The majority of such
research has used tasks that require participants to connect pieces to form complex wholes or to mentally transform objects, tasks Uttal et al. (2013) would categorize as requiring intrinsic-dynamic spatial abilities. A common spatial task given to young children requires them to recreate a design using patterned blocks. Ansari, Donlan, Thomas, Ewing, Peen, and Karmiloff-Smith (2003) found that performance on this task predicted the cardinality understanding (i.e., understanding of the purpose of counting) of 3- to 5-year-old typically-achieving children. Barnes et al. (2011) found that performance on this task predicted early numeracy skills (e.g., counting, understanding “more,” understanding cardinality) of 3-year-olds (with and without spina bifida). Farmer et al. (2013) used a similar task, where children were asked to copy two- and three-dimensional designs using shapes and connecting blocks, respectively, and found that the performance of 3-year-olds on this task was related to their basic numeracy skills at 3 years and was predictive of basic numeracy skills and math problem solving at 5 years. These same math skills were related to children’s ability to select the shape that two pieces form at 5 years. Likewise, Barnes et al. (2011) found that the ability of 5-year-olds (with and without spina bifida) to select two pieces from an array that, when put together, would create a target shape predicted their early numeracy skills including counting and object-based arithmetic skills. Some additional research has used tasks requiring participants to perceive targets amongst distracting information, tasks Uttal et al. (2013) would categorize as requiring intrinsic-static spatial abilities. Mazzocco & Myers (2003) found that the abilities of 5- to 8-year-old children to select a figure that matched a target in both shape and orientation (i.e., visual discrimination) and to select a figure that would match a target if all lines were complete (i.e., visual closure) predicted their early numeracy skills. The ability to identify a figure embedded in a design (i.e., disembedding) was related to early numeracy skills in all but the 8-year-old children. Thus, both early static and dynamic spatial skills have been linked to early numeracy skills, broadly.

To my knowledge, Gunderson, Ramirez, Beilock, and Levine (2012) are the first and only researchers to date to have explicitly examined the connection between early spatial
skills and NLE patterns. In their first study, 6- and 7-year-old children were assessed using an NLE task (0-1000 range) at the beginning and end of the school year, and were assessed on their reading ability, math knowledge, and spatial ability at the beginning of the school year. Intrinsic-dynamic spatial ability was measured using a task requiring children to choose the shape that would complete a square when rotated and put together with a target shape. It was found that when the linearity of initial NLE patterns was controlled for, spatial skill was a significant predictor of the linearity of end of year NLE patterns. Meaning that, of the children who had the same level of linearity on the NLE task at the beginning of the school year, those with greater spatial skill showed greater improvements in linearity over the year. This relationship held when general math knowledge and reading achievement were covaried; thus, it was not just a matter of generally more capable children making the greatest improvements (Gunderson et al., 2012). In their second study, children were assessed on their spatial ability at 5 years; their linearity of their NLE patterns (0-100 range) at 6 years; and their ability to perform simple symbolic and non-symbolic calculations at 8 years. Intrinsic-dynamic spatial skills were measured with the Children’s Mental Transformation Task (CMTT; Levine, Huttenlocher, Taylor, & Langrock, 1999), which requires children to choose the complete shape that two pieces would make when put together. Vocabulary was also measured as a proxy for general cognitive ability. Gunderson et al. (2012) found that linearity of NLE patterns mediated the relationship between spatial skills and symbolic calculation (but not non-symbolic calculation), and that it was not due to general cognitive ability. It is important to note that Gunderson et al. (2012) interpreted NLE patterns from a logarithmic-to-linear representation perspective, and as such they concluded that spatial skills had a positive impact on the linearity of children’s underlying mental representations of number, not simply on their performance on this task.

Additional, but indirect, evidence that spatial skills are related to NLE performance comes from training studies. Researchers have found that training aimed at improving the linearity of NLE patterns (and ostensibly underlying representations of number) of young children.

---

2 Geary et al. (2007) examined the relationship between visual spatial working memory and the NLE task in kindergarten children. While visual memory has been included as a category of spatial ability by others (e.g., Carroll, 1993), it is not included in the conceptualization of Uttal et al. (2013) used here. Indeed, while visuospatial working memory and spatial skills overlap substantially, they are not identical constructs (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001).
children was most effective when a spatial component was included (Fischer et al., 2011; Siegler & Ramani, 2009; Whyte & Bull, 2008). Siegler and Ramani (2009) found that linearity and accuracy of NLE patterns (0-10 range) were significantly more improved in low-income 4- and 5-year-olds after they spent 1 hour total playing a linear numerical board game compared to those children who played a circular numerical board game or participated in other numerical activities. Likewise, Fischer et al. (2011) found that 5- and 6-year-olds demonstrated greater improvement in the accuracy of NLE (0-10 range) when they participated in digital dance mat training compared to when they received training on a tablet PC. In the digital dance mat condition participants were presented with a target number positioned on a linear number line and a comparison number and asked to move their body to the square to the right on the dance mat if the comparison number was larger than the target, and vice versa. In the tablet PC condition, participants were asked to compare two numbers without spatial cues. All children in these studies received additional experiences with the relevant number range, but those who had training on the spatial aspects of the NLE task showed the most improvement. Whether spatial training less closely connected to the task (e.g., mental rotation training) would improve linearity of NLE patterns has not been examined; however, there is some evidence to suggest it would. Amount of video game play, a task shown to improve spatial skills in training studies (see Ferguson, 2007 and Uttal et al., 2013 for reviews), has been found to be related to linearity of NLE patterns (Ramani & Siegler, 2008). Interestingly, video game play was not associated with performance on any of the other three numerical tasks given (i.e., magnitude comparison, numeral identification, counting; Ramani & Siegler, 2008), suggesting that there is something more spatial about the NLE task than the others.

Although the evidence is limited, that which has been outlined here seems supportive of the proposal that spatial abilities are indeed relevant to NLE performance. Gunderson et al. (2012) interpreted this relationship from an underlying representation perspective. An alternative interpretation is that the NLE task does not assess an underlying representation at all, but instead, the task assesses a composite of spatial and numerical skills and that both are necessary to achieve linear, accurate pattern of estimations. From this perspective, the fact that the NLE task is robustly predictive of math achievement is explained by the fact that both spatial and numerical skills are strong predictors of math achievement themselves.
Children who have adequate skills in both of these areas would be expected to perform better on the NLE task and other math tasks, whereas children with deficits in either or both of these areas would be expected to perform more poorly.

**Visual-Motor Integration and Number Line Estimation**

In addition to numerical and spatial skills, it has been proposed that visual-motor integration (VMI), the extent to which a person can integrate their visual-perceptual abilities and fine motor abilities (i.e., hand-eye coordination), may account for some variability seen in NLE performance (e.g., Bouwmeester & Verkoeijen, 2012). The NLE task seems to require children to convert their numerical understanding into a spatial representation and then accurately mark the number line in the corresponding location (e.g., see Slusser et al., 2013). Children with weak VMI may have difficulty making this mark accurately even when they have the prerequisite numerical and spatial skills.

VMI scores can be affected by perceptual discrimination abilities, fine motor abilities, and the ability to integrate the two. A by-product of the definitional inconsistencies in the area of spatial abilities is that VMI is sometimes viewed as a spatial ability (e.g., see de Hevia, Vallar, & Girelli, 2008) despite the prominent role of fine motor skills. While perceptual discrimination is considered an intrinsic-static spatial ability by Uttal et al. (2013), VMI does not fit cleanly into any of the four spatial categories they proposed and was considered a construct separate from spatial abilities here.

To date, few studies have been conducted examining the relationship between VMI and math achievement. The research evidence that is available is mixed as to whether a relationship exists. Using a standard measure of VMI (i.e., copying geometric forms; e.g., Beery VMI), Kurdek and Sinclair (2001) found that, when age was controlled for, VMI assessed in Kindergarten was not predictive of math achievement in Grade 4. By-contrast, using the same task, Sortor and Kulp (2003) found that 7- to- 10-year-olds who were highest achieving in math differed significantly on VMI from those who were lowest achieving, even when age and verbal ability were partially controlled for (i.e., not just a matter of
intelligence). Kulp (1999) found that performance on the Beery VMI was related to teachers’ ratings of math achievement for 7-, 8-, and 9-year-olds (but not for 5- and 6-year-olds). Some additional research has used less standard, and perhaps less valid, measures of VMI. For example, Pagani and Messier (2012) found that teachers’ ratings of the VMI of Kindergarten students were related to children’s measured numerical knowledge (Number Knowledge Test; see Case & Okamoto, 1996). However, the questions asked of teachers to assess VMI related to a variety of tasks that did not purely involve VMI (i.e., related to writing and math; e.g., ability to make one-to-one correspondence, ability to classify objects by common characteristics). Using three of the large data sets used by Duncan et al. (2007), Grissmer, Grimm, Aiyer, Murrah, and Steele (2010) found that VMI at the beginning of schooling predicted later math achievement. VMI measures used ranged from the standard VMI task to tasks that were more clearly spatial in nature (i.e., copying models with blocks) and subjective ratings of VMI skills by mothers. Grissmer et al. (201) found that the strength of the relationship varied by data set. Overall, there seems to be some support that VMI is related to math ability. An important question is, why?

The findings of a study by Barnhardt, Brosting, Deland, Pham, and Vu (2005) suggest that VMI could only be superficially related to math ability. Barnhardt et al. (2005) asked 8- to 13-year-old children to copy and solve math problems. They found that children in the normal VMI group (Beery VMI >36th percentile) outperformed those in the low-VMI group (Beery VMI <16th percentile). Importantly, it was found that those in the low-VMI group showed significantly greater difficulty with alignment of numbers and organization of math problems. The authors concluded that these sorts of errors could result in incorrect answers even when the child is capable of the calculation. It must be noted, though, that many of the studies detailed above used math measures that did not require written calculations.

To my knowledge, the relationship between VMI and NLE performance has yet to be examined directly. If there is a relationship between VMI and NLE performance, it could be for one of at least two reasons. If VMI is only superficially related to math performance,

---

3 In both of these studies (Kurdek & Sinclair, 2001; Sortor and Kulp, 2003), the perceptual discrimination component was assessed separately and found to be predictive of math abilities. This is supportive of the position that VMI and spatial abilities are not synonymous and adds to the evidence that spatial abilities are robustly predictive of math abilities.
then the VMI could be seen as a confounding variable. This may not be of concern with typically developing older children who would be expected to have at least the basic VMI skills necessary for the task. However, VMI could contribute to weak performance seen in older children with VMI deficits (e.g., some children with learning disabilities) and younger children, whose VMI abilities are still rapidly developing (see Beery & Beery, 2010). Alternatively, if VMI is meaningfully related to math abilities then it could be a third aspect of the NLE task, along with numerical and spatial abilities, that contributes to its relationship to math abilities broadly.

**Malleability of Early Numerical and Spatial Skills**

In general, both early numerical and spatial skills are quite malleable when appropriate learning experiences are provided (see Levine, Gunderson, & Huttenlocher, 2011, Newcombe & Frick, 2010, and Uttal et al., 2013 for reviews). In terms of spatial skills, sustained time to engage in exploration and access to a variety of appropriate building materials and puzzles to play with (e.g., Hanline, Milton, & Phelps, 2001; Levine, Ratliff, Huttenlocker, & Cannon, 2012) appears to set the stage for young children’s learning; however, alone, these unplanned, incidental math experiences are not optimal (NAEYC & NCTM, 2010). Indeed, Casey, Andrews, Schindler, Kersh, Samper, and Copley (2008) found that when teachers presented 5- and 6-year-olds with meaningful problems to solve with blocks (e.g., build an enclosure to keep toy animals safe) and applied systematic questioning to guide them as needed, children’s performance on static- and dynamic-intrinsic spatial tasks improved significantly more than when children were provided with the same materials but no direction as to how to play with them. A similar pattern is seen in learning numerical skills. Simply increasing the time spent on numerical activities leads to improvements in abilities (Arnold, Fischer, Doctoroff, & Dobbs, 2002); however, when educators intentionally teach skills, actively engage children in discussion about and exploration of numerical concepts, and adjust the difficulty level for individual students, substantial improvements have been seen (Baroody, Eiland, & Thompson, 2009; Clements & Sarama, 2008; Young-Loveridge, 2004). The spatial and numerical growth of same-aged children, then, can be expected to vary by the quality of their learning environments.
While learning experiences have often been acknowledged as important to the development of more linear NLE patterns specifically (e.g., Barth & Paladino, 2011; Dehaene et al., 2008; Siegler & Booth, 2004), developmental trajectories have been consistently examined solely in terms of age or grade (for different number ranges) across studies (e.g., Berteletti et al., 2010; Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003; Slusser et al, 2013). This continues despite the fact that a substantial amount of variability has been seen in the NLE patterns of same-age children (e.g., Berteletti et al., 2010; Geary et al, 2008; Laski & Siegler, 2007) and that age-based developmental trends sometimes have not been seen in cohort studies (e.g., Bouwmeester & Verkoeijen, 2012; White & Szucs, 2012). When NLE performance is assumed to vary systematically by age only, other potentially impactful factors, such as spatial or numerical experience, could be erroneously treated as unsystematic variance (Bouwmeester & Verkoeijen, 2012).

There is some evidence that factors beyond age are indeed systematically related to development of linearity in NLE patterns specifically. Siegler and Mu (2008) found that most 5- and 6-year-old Chinese children showed a linear pattern of NLE performance (0-100 range) while most same-aged American children did not (62% vs. 17%, respectively). Additionally, Chinese children showed less error in estimation than American children. Siegler and Mu (2008) attributed the difference to the greater focus on math in the homes of Chinese children. Similarly, Ramani and Siegler (2008) found that 4- and 5-year-old American children from low socioeconomic status (SES) backgrounds had less linear NLE patterns than children from middle SES backgrounds. This difference was attributed partially to differences in home environment, specifically, differences in the amount of time spent playing board games at home. Children from low SES backgrounds were found to play board games at home about half as often as their middle SES peers, with the majority of low SES children never having played a board game in their home or the home of a friend or relative. The amount of board game play was found to correlate with linearity of NLE patterns of low SES children in this study. Additional intervention studies showed that as little as one hour of playing a linear numerical board game significantly improved the linearity of children’s NLE patterns (e.g., Siegler & Ramani, 2009; Whyte and Bull, 2008), and some of the gains made were maintained nine weeks later (e.g., Ramani & Siegler, 2008;
see Siegler, 2009 for a review). Thus, NLE performance seems very sensitive to environmental input, in young children at least.

What kinds of input account for changes in NLE patterns? One view is that experiences which provide multiple cues as to the order and increasing magnitude of numbers (e.g., counting to nine on your fingers takes more time and more physical space/fingers than counting to three) are most important (e.g., Ramani & Siegler, 2008; Siegler & Mu, 2008). These “redundant cues” are proposed to improve the linearity of underlying representations of numbers and thereby enhance NLE performance. An alternative viewpoint is that development in spatial skills and/or numerical skills improve strategy use on the NLE. Indeed, White and Szucs (2012) informally linked patterns of NLE performance to national curriculum standards. The youngest children in their sample appeared to count from left to right without making use of spatial cues; they noted this may have been a by-product of the expectation that they would just be learning the importance of sequences and counting. By contrast, the older children in their sample began using anchor points and were able to consider part-to-whole relationships; they noted this may have been due to the expectation that they would be learning the base-ten system, mental arithmetic strategies involving the decomposition of number, and concepts of doubles and halves. Whether the schools in their sample implemented the National curriculum with fidelity is unknown, though.

Despite clear evidence of the effects of experience on NLE performance (e.g., Ramani & Siegler, 2008; Siegler & Mu, 2008), the explicit acknowledgement of White and Szucs (2012) that instruction could have affected NLE performance of children in their samples in important ways is an exception in developmental studies of NLE. Little information beyond the age of participants, and sometimes the broad racial/ethnic and SES makeup of the sample, is typically provided in developmental studies of NLE. Given the malleability of NLE, and numerical and spatial skills more generally, comparing performance of age-cohorts within or between developmental studies without knowledge of the learning environments (e.g., curriculum content, quality of instructional techniques) of children could be misleading. Even if longitudinal studies were employed instead of the typical cohort studies (as Bouwmeester & Verkoeijen, 2012 recommend), details about the learning environments
children experience should be detailed if generalizations about the rate of development were going to be made.

**Current Study**

In the current study, the NLE (0-10 range) performance, numerical knowledge, spatial skills, and VMI of children learning in an inquiry- and play-based Junior Kindergarten classroom were examined near the beginning and end of the academic year. There were a number of advantages of the design used in this study over previous studies. First, at both time points, all numbers between the anchors on the NLE task (i.e., 1 through 9) were presented as target numbers twice (see also Whyte & Bull, 2008), as oversampling the lower range has been purported to bias model selection in favour of a logarithmic-to-linear shift account (Barth & Paladino, 2011; Barth et al., 2011). Second, the NLE task was given at two time points, allowing development in NLE to be examined longitudinally, rather than across cohorts (Bouwmeester & Verkoeijen, 2012). Gunderson et al. (2012) also looked at changes in NLE over the school year, but a linear-to-logarithmic shift account was assumed. A third advantage of this study was that the classroom environment between the time points was documented through classroom observations and an interview with the classroom teacher. This allowed the effects of instruction on development in NLE patterns and numerical, spatial, and VMI abilities, to be considered in addition to the effects of maturation. There were three main objectives of this study.

One main objective of this study was to determine whether a strategy account (White & Szucs, 2012) may better explain NLE performance than a logarithmic-to-linear shift account (e.g., Opfer et al., 2011) or a proportion-judgment account (e.g., Barth & Paladino, 2011). Evidence for and against the logarithmic-to-linear and proportion-judgment accounts exists in the literature (e.g., Ashcraft & Moore, 2012; Barth & Paladino, 2011; Bouwmeester & Verkoeijen, 2012; Núñez, 2011; Opfer et al., 2011; Slusser et al., 2013) but a strategy account was only recently, and fairly informally, suggested (see White & Szucs, 2012). Given that linearity of NLE performance has been linked to math achievement (Booth & Siegler, 2006; Siegler & Booth, 2004) and math learning disabilities (Geary et al., 2007, 2008), having an accurate understanding of what contributes to individual differences in
performance is vital to creating appropriate interventions. The following criteria were used to establish which of the three accounts was best supported by the data in the current study.

For a logarithmic-to-linear shift account to be supported, the majority of children’s individual NLE patterns at both times would be best fit by a logarithmic or linear function rather than an unbounded power, one-cycle power, or two-cycle power function. If the shape of children’s NLE patterns changed from the beginning to the end of the study, the change would involve a progression towards greater linearity. That is, a shift from a linear to a logarithmic fit or a drop $R^2$ values would not be expected. Additionally, if children were relying solely on an internal representation of number to complete this task and the patterns seen were aligned with this representation, it would be expected that placement of the same target numbers on the number line would be consistent from one trial to the next only minutes later. Large differences in target estimations would suggest that something other than, or in addition to, an internal representation of target numbers was being assessed by the NLE task.

For a proportion-judgment account to be supported, the majority of children’s individual NLE patterns at both time points would be best fit by an unbounded power, one-cycle power, or two-cycle power function rather than a logarithmic or linear function. If the shape of children’s NLE patterns changed from the beginning to the end of the study, a progression towards linearity would be expected. That is, a shift from a two-cycle power function to a one-cycle power function (or from a one-cycle power function to an unbounded power function) or $\beta$ values of the best-fitting function moving further away from 1 would not be expected. For evidence of strategy use (i.e., patterns of absolute error and variability in error for individual numbers estimated) to be consistent with a proportion-judgment account, the tent-shaped and M-shaped error patterns, described by Ashcraft and Moore (2012) and also observed by White and Szucs (2012), would correspond with NLE patterns best fit by one-cycle and two-cycle models respectively.

A strategy account would be considered only if neither the linear-to-logarithmic nor the proportion-judgment were supported but evidence of strategy use existed. For a strategy account to be supported, sophistication of strategy use (i.e., greater use of anchors and imagined reference points) would be related to linearity of NLE patterns. Also, if strategy
use changed from the beginning to the end of the study, strategy use would become more sophisticated. That is, a shift from an M-shaped error pattern to a tent-shaped error pattern would not be expected. White and Szucs (2012) examined the strategy use in relation to grade of children and then linked strategy use to math learning by grade, informally. I have extended the statistical analysis of White and Szucs (2012) to an examination of the relationship of strategy use with measured spatial and VMI abilities. Children with better developed spatial skills and VMI may be more capable of converting their understanding of the part-whole relationships of numbers to an accurate spatial representation of this relationship on the number line.

A second main objective of the study was to look at the relationships between numerical knowledge, spatial skills, and NLE. Specifically, it was expected that linearity of NLE patterns would be related to numerical knowledge, at both time points, regardless of which account was best supported because this relationship was predicted by all accounts. Additionally, error patterns, which are indicative of strategy use, were expected to be related to numerical knowledge; children with greater numerical knowledge would be expected to use more sophisticated strategies. In line with past research (e.g., Barnes et al., 2011; Farmer et al., 2013; Mix & Cheng, 2012) it was predicted that spatial ability would be related to numerical knowledge in this sample. It was also expected that the results of Gunderson et al. (2012) would be replicated in this slightly younger sample; it was predicted that spatial ability would related to linearity of NLE patterns and that children with greater spatial abilities would show greater improvements in linearity.

A third main objective of the current study was to investigate whether VMI was a confound in the NLE task or a correlate that increased its value in predicting other math abilities. Poor VMI could affect performance on the NLE task by affecting the accuracy of placing a mark in the intended place on the number line. Previous research has shown mixed evidence with regards to the relationship between VMI and math ability (Kurdek & Sinclair, 2001 vs. Sortor & Kulp, 2003) and the relationship between VMI and NLE performance has yet to be examined. The relationship between VMI and linearity of NLE performance and VMI and numerical knowledge were examined at both time points. VMI would be considered to have added value in the prediction of math abilities if it were related to both
numerical knowledge and linearity of NLE patterns. VMI would be considered a possible confound of the NLE task if it were related to linearity of NLE patterns but not numerical knowledge.

A secondary objective of this study was to describe an early instructional environment that meets the standards set out by the NCTM (2000) and NAEYC (NCTM & NAEYC, 2010) and to track the development of early numerical and spatial skills of the learners in this classroom. Numerical and spatial skills have been shown to be predictive of later math achievement (e.g., Aunola et al., 2004; Cheng & Mix, 2012; Duncan et al., 2007; Gunderson et al., 2012; Jordan et al., 2010; Mix & Cheng, 2012) and, importantly, quite amenable to change (e.g., Casey et al., 2008; Levine et al., 2011; Siegler, 2009; Uttal et al., 2013). Concrete examples of best-practices in math education in the early years and demonstrations of how much young children are capable of spatially and numerically under the right circumstances may be useful resources to educators who are aiming to improve their math teaching. The NCTM and NAEYC (2010) explicitly acknowledge that such resources are needed to support educators and to improve early childhood math education. In the current study, comparisons between performance of this relatively unique sample and population averages were made on all measures where normative data was available. Additionally, a comparison of average scores on each measure at the beginning and the end of the study was made. It was expected that significant growth would occur on all measures.

Methods

Participants

Children were recruited from a Junior Kindergarten class in a laboratory school in Ontario, Canada, following approval from the research ethics boards of both the University of Toronto and the school. Only one parent did not consent to their child participating in individual testing. The right of the children to choose not to participate was respected; at the first time point one child chose not to complete two measures (Number Knowledge Test-Abridged and Number Line Estimation Task).

Twenty-one 4-year-old children (10 boys, 11 girls) participated in the study. At the beginning of the study, their mean age was 52 months (SD = 2.90; range = 47 to 57 months).
Information on race, ethnicity, and socioeconomic status was not collected; however, it was expected that the diversity of the sample was in line with the diversity of the laboratory school. All children were residents of a culturally diverse, metropolitan city with 44% of the school’s current students self-identifying as visible minorities. About 12% of the school’s current students received some financial assistance with respect to tuition; however, all families with children enrolled in the Junior Kindergarten classroom were capable of paying the full tuition fee ($16,448 plus a $5000 enrolment fee for the 2012/2013 academic year) for one or more children independently. Many parents place their children on a waiting list for admission to the laboratory school soon after their birth. Children are not admitted based on academic merit and about 15% of the school’s current students receive some special education support.

Classroom

Children were in the classroom all day Monday to Friday, except for Wednesday afternoons, when teachers engaged in professional development. The classroom was staffed by a teacher, interns, and an assistant. The classroom teacher had been a teacher for 16 years, teaching pre-kindergarten to Grade 2 in both public and private schools. She held a MA in Philosophy, a MA in Child Study and Education (with qualifications to teach), and a PhD in Developmental Psychology. Interns were either in the first or second year of their MA in Child Study and Education at the university associated with the laboratory school. A first year intern assisted in the classroom four mornings each week and a second year intern assisted five mornings and two afternoons each week. A paid assistant, a university student from the work-study program, assisted at least two afternoons each week. This means that most mornings there were at least three adults in the classroom and most afternoons there were two adults.

The classroom was quite small. Half of the classroom was shared with the pre-kindergarten classroom next door and was only available for part of the day. There was little technology; there was only one computer and no SMART Board™, which are pervasive in Ontario classrooms. The classroom contained the centers seen in typical kindergarten classrooms: house center, dress-up center, block center, painting center, sandbox, and book corner. A wide variety of materials intentionally selected to stimulate mathematical thinking
were available (see Figures 2 and 3). Some criteria were used by the classroom teacher in selecting materials. The teacher aimed to select materials that were open-ended and could be used and combined in a variety of ways or materials that had specific challenges built into them, but these goals could be achieved in a variety of ways. The teacher noted that she liked to have a variety of materials that were proportional and regular (e.g., one large block is equivalent to two medium blocks or four small blocks) because children pick up on and exploit this proportionality, as well as a variety of materials that were aesthetically pleasing (i.e., have weight to them, have a variety of textures), because one goal of children in play is to create beautiful things. An assortment of surfaces (e.g., frames, peg boards, light table, grids; see Figures 2a, 2e, 2j, 2l, 3a, 3d, 3f, 3g, and 3h) on which to work with materials were also provided. The teacher noted that she provided these to create some structure and organization in children’s play that helps mathematical thinking come about. She believed that a balance of freedom and boundaries are required for productive play.

The classroom teacher described the classroom as having a child-centered, play-based program. About half of each day was devoted to sustained play and about half was devoted to teacher-directed instruction. Play included free play times, where all centers were open and many materials were available, and structured building and puzzle play times, where children were provided with a limited number of teacher selected materials. During structured play times the teacher selected materials because they had similar properties she wanted children to explore (e.g., fit together) or because they would encourage them to build in new ways (e.g., providing a small platform may encourage them to build up rather than out). She would also sometimes set problems for children to solve during these play times. These play periods helped assure that children who may not have naturally gravitated towards these materials had opportunities to experience using them and helped all children become more organized in their play. The teacher pointed out that materials selected do not serve their intended mathematical purpose when used randomly and chaotically (e.g., using small blocks and gems as “food” to cook in the house center). When children become more organized in their play (e.g., placing items with purpose rather than dumping items in piles), the potential for patterning, counting, and thinking about symmetry arises.
While children were playing, the teacher and the interns were watching and listening. They posed questions or made comments to highlight or extend children’s mathematical thinking. For example, overhearing a child working with beads on a peg-board note, “It’s almost a square, but not quite.” the teacher asked, “If you wanted to make it a square what would you do?” The staff were careful to not interrupt or overtake play by asking questions that did not coincide with the purposes and interests of the children (e.g., asking them to count the number of blocks around the perimeter of the house they are building). Often, the staff would plant a seed (e.g., a brief comment and gesture about the symmetry of a design) and then allow children to do with it what they pleased. Children were also subtly encouraged to sustain their focus on one activity rather than flitting from one activity to the next. For example, a staff member may have simply shown interest in what children were working on (e.g., taking a picture with the classroom camera), placed additional materials nearby that may extend the play, posed a challenge (e.g., “I wonder if…”), or sat and provided some scaffolding or assistance when a task was becoming difficult. With these subtle encouragers, children were observed to sustain focus on one activity for upwards of an hour near the end of the study. Following, some specific examples of children’s play, documented through observations and an interview with the classroom teacher, are presented.

**Sustained Play: Example 1.** In this example, the teacher and the intern helped children sustain and organize their play by showing interest and by providing challenges that were consistent with the children’s own interests and goals. Math ideas and language (e.g., “equal”) were brought to the fore. Children can be seen engaging in joint-problem solving and beginning to learn from one another.

During free playtime, two boys were working together to build a tower using a large bin of Wedgit™ building blocks (see Figure 4a). Although the Wedgit™ tower was several inches smaller than either boy, one boy excitedly exclaimed, “It’s even taller than us!” The other boy called the teacher over to show her their tower. She showed great interest and challenged them to try to make it as tall as one of the boys. An intern used the class camera to take a picture of the tower. Another child was drawn into the excitement, and offered, “If you put that in it will make it taller.” The child was referring to a small cube that would have fallen through the existing hole and would not
have added to the height. One of the boys indicated that it would not work and the child asked “Why not?” Before an explanation could be given, the tower fell. Without missing a step, they all started rebuilding. The teacher offered a hint about how to use the blocks; she showed how one block turned vertically could be wedged into another block sitting horizontally: “Did you know they can go the other way?” The boys continued building, but did not take this advice. The teacher allowed them to go about what they were doing. They took turns seeing if anyone was shorter than their tower, yet. After 25 minutes, two boys returning from the library joined in the building. One child began to build his own tower. The group, up to five children now, was getting quite excited. The teacher encouraged them to focus and cooperate by indicating, “We are going to try to make them equal.” After nearly an hour since they began building their tower, the boys notice the teacher and the interns helping other children tidy up their areas and realized it was time for snack. They started tearing down their tower and putting the blocks away.

**Sustained Play: Example 2.** In this example, the teacher encouraged sustained, purposeful play by: lending a helping hand when the task became frustrating; encouraging the use of additional materials; and conveying that their play was important by moving the next class activity to another room. The children applied a concept (symmetry) that they were exploring as a class and the teacher reflected this back to them.

Spotting two girls wandering aimlessly during free play time, the teacher simply suggested “try building with those blocks.” The girls decided to build a sidewalk (see Figure 4b). The girls added a curve to one side of the sidewalk and searched for an identical curve to make the design symmetrical. The teacher noticed the symmetry and verbalized what the children were doing. She helped the girls try to find the matching curve. They concluded that all the curves were a little different from one another. One of the girls took a short curve and added a rectangular block to it to make it match the other side. They added additional wooden cylinders and coloured blocks, maintaining the symmetry. One of the girls held up her hands to make a frame and commented that she was taking pictures. She shared that when she stood back further, she could see more of the picture. When the teacher noticed the girls were beginning to become
somewhat bored with the activity, she encouraged them by saying “You need some people. People deserve to live in this beautiful place.” The girls went and retrieved baskets of miniature trees, animals, insects, and people and began adding them to their creation. Since they were once again showing excitement for the task, the teacher arranged to move their ‘reading buddies’ to another room for the next period, so the girls could continue building later. After reading buddies, the girls excitedly invited another girl to see what they had done and to build some more with them.

The teacher also took note of mathematical ideas or interesting examples of mathematical concepts that arose in play to bring into periods of teacher-directed math instruction. The teacher used these examples as an in-road to the lesson where she and the children reflected on the ideas, learned the related vocabulary, and extended the ideas to set new challenges for learning. Explicit math instruction took place in half-class groups (while the other half of the class was in the library or at French class) at least twice per week (for each group). The teacher also worked with smaller groups and individual children based on skill level and needs, and incorporated math concepts into whole group activities (e.g., counting tallies in a game, figuring out how many children were missing). The teacher shared that she has a standard set of topics in math that she likes to explore with her class each year, but she noted that the emphasis on and depth to which she went into topics was dependent on the group. Some of the math topics that were explored between December and May are listed in Table 1. Children participated in activities that were carefully constructed to give them opportunities to investigate these math ideas and practice these math skills while still being quite playful. Some specific examples of teacher-directed instruction follow.

**Teacher-Directed Instruction: Example 1.** In this example, the teacher has intentionally introduced a new math concept and used this opportunity to practice some basic numerical skills (e.g., counting, determining “more” and “less”) as well as some basic literacy skills (i.e., reading and writing fairly familiar words). This example illustrates the use of a game-like context to enhance children’s interest in math. The teacher reported that dice were frequently used in math activities (e.g., for Snakes-and-Ladders-like games on 1 to 25 boards and the school stairs) for this reason.
Half of the class had just left for French class. The teacher had the remaining children count how many children are on the carpet. Together, they read the names of the children who were away and those that had left for French. Then, one by one, the teacher pulled out graphs made by the other half of the class showing: how many children had birthdays in each month of the year; how many boys and girls there were; and how many children were 4 years old or 5 years old. For each graph, the teacher asked questions such as “Are there more boys or girls? How do you know?” and, while pointing to a bar on the birthday graph, “What does this mean?” They counted the name blocks that made up the bars together. The teacher pointed to each block as she counted, highlighting one-to-one correspondence. The children then split into three groups. Each group went to a table with the teacher or one of the two interns to add their name blocks to the right places on the graphs. For two graphs, the children found their names pre-printed on the table and taped them on to the graph themselves and for the other graph the children printed their names and drew small pictures of themselves to place on to the graph. Some children needed help printing their own names; for these children, the teacher described the letters to the child in terms of the lines and shapes needed to form them. For each graph, the teacher and the interns asked the children questions, such as why they put their name block directly above the previous name rather than leaving a gap in between (i.e., highlighting the fact that the unit of measure is important). Children moved to the next graph whenever they were ready, rather than traveling in groups. When children were done with all three graphs, they moved to the carpet for free play.

Additional graphing activities took place in the days and weeks following. In one activity, children rolled a die and graphed how many times each number appeared. The task was presented as a game where the numbers were racing and the children had to find the winner. Many children moved on to using two dice and graphing the sum of the dice (see Figure 5). They were amazed to find that the number 7 was the winner for all of them. The teacher capitalized on their interest and helped them figure out why; she suggested that they find all the ways that each number could be made. By completing this task, the children discovered that 7 could be made the most ways and that is why it always won. Some children also pointed out that their teacher had
“tricked” them by putting the number 1 on the graph because there was no way to make 1 with two dice. Some children moved on to graphing the sums of three dice and some children were excited to take the game home so they could keep playing.

Teacher-Directed Instruction: Example 2. In this example, the teacher provided several opportunities for children to practice basic numerical skills (e.g., numeral recognition, counting). At the same time, she has provided an opportunity to the children who were ready for it to work with groups of five, solve meaningful arithmetic problems, and explain their reasoning. Once again, a game-like context was used.

Half of the class had just left for French class. The teacher had the remaining children stand around the outside of the carpet. She held up numbers and had them perform actions (e.g., clapping over head, hopping on one foot) that many times, cleverly ending with sitting on the carpet for the number one. The teacher then brought out a 100s chart with all of the numbers removed. She gave each child a number card (1-10) that they had to keep a secret. Working in order from 1 to 10, children came up and placed their secret number in the correct place on the chart. Before working with the numbers 11 to 20 in the same way, the teacher had the children predict what they thought the last number in the next row would be and explain why. They counted the empty spaces together to check their predictions. Once all the numbers 1-20 were up on the chart, the teacher introduced a game. She would point to the numbers for them to count, sometimes skipping a number or moving backwards (e.g., 1, 2, 3, 4, 3). If she was able to “trick” them, she earned a point, otherwise they did. She kept track of the points using a tally system. She explained how the tally system worked (i.e., one tally for each point and make a group of tallies at five) and thought aloud each time she put another point up (e.g., “That makes one, two, three, four. That means I will have to put one across next time to make five.”). She also asked questions related to who was winning as she went (e.g., “How many more do I need to catch you?”). At the end, they counted the tallies together. The teacher modeled a count-on strategy, starting with the group of five and counting six, seven, and eight. The children were excited because they recognized right away that they won. The teacher challenged them, though: “I don’t think you won because I think 3 is bigger than 8.” In unison they
disagreed and the teacher asked, “How do you know?” A few children were given an opportunity to explain their thinking.

Sometimes the classroom teacher would sit down with parents and discuss specific things they could do with their individual child at home to help them progress to the next step (e.g., sorting silverware, playing board games) but homework was rarely given. There was an assumption that for this population of children, sufficient informal and formal learning opportunities are happening outside of the classroom without teacher direction.

**Measures**

**Number Line Estimation Task.** A 0-10 number line was used for the NLE task. The numbers 1, 2, 3, 4, 5, 6, 7, 8, and 9 were each presented twice, totaling 18 trials (similar to Whyte & Bull, 2008). Additionally, because the children were so young and may have had little formal experience with number lines, they were presented with an orienting item and a training item prior to beginning the task. The training item did not involve using the midpoint of the task range, as some researchers have (e.g., Barth & Paladino, 2011; Booth & Siegler, 2006), because this has been criticized as teaching a strategy (i.e., using the midpoint; Opfer et al., 2011). Instead, similar to Ashcraft and Moore (2012) and White and Szucs (2012), a range half the size of the task range was used for the orienting and training items (i.e., their smallest task range was 1-20 and their training item was 0-10).

To begin, children were told: “We are going to play a game with number lines.” and then asked, “Do you know what a number line is?” Regardless of their answer, children were shown the orienting item, a complete 0-5 number line (Figure 6a). The number line was 9.25 cm long (half as long as the 0-10 number line) and was printed in black ink on a 10 cm x 28 cm white page, beginning 4.25 cm from the left side. A 0 was printed above a 1.4 cm perpendicular line anchoring the left side and a 5 anchored the right side in the same way. Hash marks 1.9 cm long were drawn perpendicular to the horizontal line (extending 0.5 cm below it) at equal, 1.8 cm intervals between the anchors. The numbers 1, 2, 3, and 4 were printed directly above the hash marks. The font for all numbers was selected carefully to ensure that numerals were in a format most familiar to children (e.g., 4 vs. 4). A 20 pt. font was used. Children were told, “This is a number line. This number line goes from 0 to 5
with 0 on this end and 5 on this end. The numbers 1, 2, 3, and 4 go in order in between and have the same amount of space between them. There is a mark through the number line showing just where each number belongs.” The examiner pointed to the anchors and to the equal spaces when they were named.

Children were then shown a training item, an empty 0-5 number line of the same size and orientation with the number 4 (28 pt. font) printed inside a 2.2 cm x 1.2 cm box, 3 cm from the right side of the page (Figure 6b). Children were asked: “If this is 0 and this is 5, where does 4 belong on this number line? Make a mark where 4 would go.” The examiner pointed to the anchors when they were named. If children did not make a hash mark all the way through the horizontal line, the examiner demonstrated how to do so. If children responded somewhere on the left-hand side of the number line, the examiner brought out the orienting item again and said: “Let’s look at the number line and check if 4 is where it belongs. Hmm. It looks like the 4 should be a little closer to the 5. What do you think?” The examiner did not continue until it was clear children understood the goal of the task. No strategies for completing the task were given. Once children understood the task, the examiner said: “Great! Let’s try some more.”

For each trial of the task, children were presented with an empty 0-10 number line (Figure 6c). The number line was 18.5 cm long and was printed in black ink on a 10 cm x 28 cm white page, beginning 4.25 cm from the left side. A 0 was printed above a 1.4 cm perpendicular line to anchor the left end and 10 anchored the right end. Each target number was printed in a 28 pt. font inside a 2.2 cm x 1.2 cm box, above and 1 cm to the right of the number line. The number was not printed above the middle of the empty number line as other researchers have done (e.g., Booth & Siegler, 2006; Siegler & Opfer, 2003) because it was observed during piloting of the task that some children had a tendency to draw the hash mark directly under the target number, regardless of its magnitude. A similar tendency was seen in children in other studies (e.g., see Figures 3 and 5 in Barth & Paladino, 2011). Whether these children misunderstood the task or simply responded in this fashion because they did not have the knowledge to complete the task is unknown. By moving the target number away from the number line it was thought that the likelihood of misunderstanding would be reduced.
On the first trial of the task, children were told: “This number line is different. It goes from 0 to 10, with 0 here and 10 way down here. There are lots of numbers that belong between 0 and 10.” The examiner pointed to the anchors when they were named. The examiner then asked: “Can you count from 0 to 10?” Children were given positive feedback for their counting (their performance was not recorded) and then the examiner continued, “I am going to ask you to show me where some of those numbers belong on the number line. Okay?” For each target number (N) the examiner asked: “If this is 0 and this is 10 where does N belong?” Anchors were pointed to when they were named. Children were presented each of the target numbers 1 through 9 once before being shown any target number a second time. The numbers were presented in the same order to all children at both time points: 1, 5, 2, 9, 4, 6, 3, 8, 7, 1, 6, 2, 5, 8, 4, 9, 7, 3. Effort was praised, but no feedback as to accuracy of responses was provided.

**Number Knowledge Test (Abridged).** The Number Knowledge Test (NKT) was developed by Case and Okamoto (1996) to align with their idea of a central conceptual structure for number (i.e., a “mental counting line”). It is a reliable test (Case & Okomoto, 1996) that has been used as a progress monitoring tool in intervention studies (Griffin, 2004) and has been described as the best screening measure for future math achievement (Gersten et al., 2005). Case and Okamoto adapted the test for the use in the National Longitudinal Survey for Children and Youth (NLSCY) a long-term study tracking the development of multiple cohorts of tens of thousands of Canadian children from birth through to adulthood. The test was shortened to be used with only 4- and 5-year-olds and a new discontinue rule was put in place (i.e., stop after three consecutive incorrect items). This adapted version of the NKT will be referred to as the Number Knowledge Test (Abridged; NKT-A). The NKT-A has been used in the NLSCY since 2000-2001 (Cycle 4). In an evaluation of the NKT-A using NLSCY data, it was concluded that the NKT-A is a valid measure of the numerical knowledge of young children (Statistics Canada, 2004). This abridged version was used in the current study to allow the sample to be compared to available Ontario normative data.

The NKT-A (Case & Okamoto, 1996; Statistics Canada, 2004) contains 22 items tapping numerical understanding within three developmental levels.
**Level 0 – Pre-dimensional Level.** Children at this level are thought to be able to count sets accurately (i.e., understand stable order, one-to-one correspondence, order irrelevance, cardinality, and the abstraction principle; Gelman & Gallistel, 1978) and understand relative quantity at a global level (e.g., compare quantities of concrete objects; Case & Okamoto, 1996). These skills are not well integrated, however. This level is assessed using five items involving counting chips to determine “how many” and determining which pile of chips has more or less. A child must get 3/5 correct to pass this level.

**Level 1 – Uni-dimensional Level.** At this level, children’s skills for determining exact quantities and basic understanding of relative quantity are thought to be integrated to form the central conceptual structure of a “mental counting line.” Children are able to use this line to solve problems involving exact, symbolic values (Case & Okamoto, 1996). There are eight items at this level involving counting forward and backward in the 1-10 range; determining which number is bigger or smaller in the 1-10 range; and mental addition word problems involving one-digit numbers (i.e., 2 + 1, 4 + 3). A child must get 5/8 correct to pass this level.

**Level 2 – Bi-dimensional Level.** At this level, children are thought to be able to use two mental counting lines simultaneously allowing them to work with the “tens” and “ones” that make up our base-ten number system (Case & Okamoto, 1996). There are nine items at this level involving: counting forward and backward in the 1-100 range; determining which number is bigger or smaller in the 1-100 range; determining how many numbers fall between two numbers in the 0-10 range; mental addition and subtraction problems involving two-digit numbers (i.e., 12 + 54, 47 – 21); and providing the largest and smallest possible two-digit numbers. A child must get 6/9 to pass this level.

Seven of the 22 items have a Part A and a Part B. A child must get both parts correct in order to score a point. This is to account for guessing on items with a two-choice response format. Two scores were calculated: i) a raw score (1 point per item to a maximum 22); and ii) an age-equivalent (AE) score (maximum 3). An AE score is indicative of how many levels a child passed (e.g., a child with an AE score of two would have passed the Level 1,
the uni-dimensional level). Case & Okamoto (1996) expected children to pass these levels at ages 4, 6, and 8 years, respectively.

**Children’s Mental Transformation Test.** The Children’s Mental Transformation Test (CMTT; Levine et al., 1999) is a widely used (e.g., Ehrlich, Levine, & Goldin-Meadow, 2006; Farmer at al., 2013; Gunderson et al., 2012; Harris, Newcombe, & Hirsh-Pasek, 2013; Levine et al., 2012) measure of intrinsic-dynamic spatial skills designed for young children (4- to 7-year-olds). For this study, 16 items targeting mental rotation were selected from the 32-item CMTT, Form D (Order 1). Each item involves a different target figure. The target figures are split vertically or horizontally into two symmetrical pieces. These pieces are separated and rotated 60 degrees away from the vertical axis. For half of the items, the pieces were horizontally aligned 2 cm apart at their closest points. For the other half of the items, the pieces were diagonally aligned, their closest points separated by about 2.83 cm diagonally (that is, separated by 2 cm both vertically and horizontally; see Levine et al., 1999, Figure 2). The 16 items that were not used required only translation (i.e., moving the pieces together without rotating them). The pieces that make up the target items are typically printed on one card and the four response options (all symmetrical figures, one of which is the completed target figure) are arranged in a 2 x 2 formation on a separate card. The target card is typically placed above the response card (see Levine et al., 1999, Figure 1). In the current study, for ease of administration, the items were printed in landscape orientation on 8.5 in x 11 in white paper and bound into a book, such that when the pages were turned the target pieces were automatically above the response options. Identical to Levine et al. (1999, p. 942), on the first item the examiner said “Look at these pieces. Look at these pictures. If you put these pieces together they will make one of the pictures. Point to the picture these pieces make.” The examiner pointed to the pieces and pictures when they were named. On all following items, the examiner said, “Point to the pictures these pieces make.” All items were administered to all children, regardless of performance. One point was scored for each correct answer to a maximum of 16 points. Most (71%) children performed above chance (chance = 4 out of 16) at Time 1 and all children performed above chance at Time 2.

**Beery VMI (Short form).** The Sixth Edition of the Beery-Buktenica Developmental Test of Visual Motor Integration (better known as the Beery VMI) is reliable, valid, and
widely used standardized measure of VMI (Beery & Beery, 2010). In this task children are asked to copy increasingly complex geometric forms. For young children, the first three items are drawn by the examiner and then copied by the child. The following items are copied from printed line drawings. The full-form contains 24 printed forms (plus 6 prerequisite items) and was normed for people 2 through 100 years old. The short-form is identical, but only contains the first 15 forms. It is considered adequate when assessing children up to 8 years old. The test was administered and scored in accordance with the Administration, Scoring, and Teaching Manual (Beery & Beery, 2010). Children could earn a maximum raw score of 21; corresponding standard scores with a mean of 100 and a standard deviation of 15 were also calculated. It should be noted that at the first time point two children did not meet the criteria to discontinue (i.e., three consecutive missed forms) before they had completed the entire booklet. Four children did not discontinue at the second time point. None of these children earned a perfect score (all scored zero on two or more forms); it is possible, but unlikely, that they would have earned additional points if the full-form had been used.

**Procedure**

The intention was to administer the tasks at the beginning of the school year (i.e., September/October) and the end of the school year (i.e., May/June) as Gunderson et al. (2012) did. Although I spent some time playing alongside the children in their classroom in September to allow them to become familiar with me, they were quite hesitant to participate in the testing in early October. The original testing rooms were located in the basement of the school, below the children’s classroom and children were asked to come with myself and the other examiner individually. In consultation with the classroom teacher, it was decided that it was best to wait until December to begin administering tests to allow the children to become more settled in. I spent some additional time in the children’s classroom, as well.

In early December, children seemed much more comfortable participating. They were taken in pairs to larger, brighter offices. One child from the pair completed the NLE and NKT-A while the other child completed the Beery VMI and the CMTT, then the children were accompanied back to their classroom together. They came for a second session, during the same week, to complete the other set of tasks. The NKT-A and NLE were administered
by a male research officer who sat beside the children, to the right during each task. The Beery VMI and CMTT were administered by the author. I sat directly across from the children during each task. All four tasks took 20-30 minutes to administer. The Beery VMI and the CMTT were always administered together with the Beery VMI administered first. The NKT-A and NLE were always administered together with the NKT-A administered first. The order in which the children completed these pairs of tasks was random. The same procedure was followed 5 months later, in May.

**Data Analysis**

Children estimated the position of target numbers by drawing a hash mark through the number line. Each of these hash marks was converted into a numerical estimate of the target number, employing the same method used by Siegler and colleagues (e.g., Booth & Siegler, 2006; Siegler & Opfer, 2003) and those that followed (e.g., Barth & Paladino, 2011; Berteletti et al., 2010; White & Szucs, 2012). First, the distance from the left anchor to the hash mark was measured. Then, this measurement was entered into the following general equation:

$$\left( \frac{\text{distance from the left anchor (mm)}}{\text{total length of number line}} \right) \times \text{scale of the number line}$$

In this study, the total length of the number line was 185 mm and the scale of the number line was 10 (i.e., 0-10 range). Therefore, the following specific equation was used:

$$\left( \frac{\text{distance from the left anchor (mm)}}{185} \right) \times 10$$

So, for example, if a child drew a hash mark 140 mm from the left anchor for the target number 9, the target number estimate was in the position of 7.57.

These target number estimates were used in three main ways: 1) plotting estimation patterns to determine models of best fit; 2) calculating a measure of linearity to be used in other analyses; and 3) plotting patterns of error in estimation for individual numbers to look for evidence of strategy use. These analyses were done at both the individual child and the group level at both time points.
Additionally, the consistency of children’s estimations of target numbers between trials at the same time point was examined. For individual children, the percent absolute difference in estimation (PADE) of each target number was calculated using the following equation:

\[
\text{PADE} = \left| \frac{\text{target number estimate trial 1} - \text{target number estimate trial 2}}{\text{scale of the number line}} \right| \times 100
\]

Then, the mean PADE was calculated across all nine target numbers. A child with a mean PADE over 20% was considered to be inconsistent in his or her responding.

**Model Selection.** NLE patterns were analyzed to determine which model best described the data. Three models predicted from the proportion judgement account (i.e., unbounded power, one-cycle power, two-cycle power; Barth & Paladino, 2011; Slusser et al., 2013) and two models predicted from the logarithmic-to-linear shift account (i.e., logarithmic, linear; e.g., Siegler & Opfer, 2003; Opfer et al., 2011) were examined. The scatters of target number estimates plotted against target numbers were first examined for bivariate outliers. Outliers were considered to be points that were substantially off-trend visually. These were removed before comparing models. Models were compared using the Akaike information criterion with a correction for small sample size (AICc) as is recommended by Barth and colleagues (e.g., Barth & Paladino, 2011; Slusser et al., 2013) when models involved different numbers of parameters. Typically, the best-fitting model was considered to be that with the lowest AICc value. However, when the lowest AICc value differed from other AICc values by less than two points, these models were considered to be equally good fits statistically (Burnham & Anderson, 2002). In these cases, a judgement was made based on the best fit visually. When one model was not visually a clear choice, the most parsimonious model was selected (e.g., linear; see Browne, 2000). \(R^2\) values were also calculated and reported because of greater familiarity with the meaning of this statistic amongst most readers (Slusser et al., 2013). A Microsoft Excel worksheet available freely online (Slusser & Barth, n.d.) was used to determine these values for all five models. Models were fitted to individual estimation patterns at both time points as well as group median estimation patterns at both time points. \(\beta\) values of best-fitting power functions were also calculated for those NLE patterns that were best fit by a power function.
**Linearity of Estimation Patterns.** Children with more linear estimation patterns are often assumed to have a better understanding of the linearity of a number line, that is, an understanding that numbers are equally spaced in ascending order. Two measures of linearity of estimation patterns were noted in the introduction: $R^2_{lin}$ (e.g., Booth & Siegler, 2006; Siegler & Booth, 2004) and the $\beta$ values of best-fitting power function (Barth & Paladino, 2011; Hollands & Dyre, 2000). Shifts in these values over time were examined for individual children to see if the shift was consistent with what would be predicted by the linear-to-logarithmic shift account and the proportion-judgment account, respectively.

$R^2_{lin}$ is the most common measure of linearity, but this statistic is really only appropriate when the pattern seen is best fit by a line and it can be misleading when applied to the NLE task if slope and $y$-intercept values are not taken into consideration (Bouwmeester & Verkoeijen, 2012). For example, a child whose estimation pattern is best fit by the line $y = x$ and a child whose estimation pattern who is best fit by the line $y = 0.25x + 1$ could both have $R^2_{lin} = 0.9$ (see Figure 7). In this case, one child consistently estimates accurately while the other consistently becomes more inaccurate. Looking at deviations from a perfect estimation ($y = x$) may give a better indication of children’s understanding of the linearity of the number line (Geary, 2011). In this study, mean percent absolute error (mean PAE) in estimating target numbers was used as a measure of children’s deviation from a perfectly linear pattern of estimation (i.e., $R^2_{lin} = 1$, slope = 1, $y$-intercept = 0). PAE was calculated for each trial of each target number employing the same general equation used by Siegler and colleagues (e.g., Booth & Siegler, 2006; Siegler & Booth, 2004) and others (e.g., Slusser et al., 2013; White & Szucs, 2012).

$$\left|\frac{\text{target number estimate} - \text{target number}}{\text{scale of the number line}}\right| \times 100$$

Continuing with the above example, if a child estimated the target number 9 to be in the position of 7.57, the PAE of the estimate would be:

$$\left|\frac{7.57 - 9}{10}\right| \times 100 = 14.3$$
That is, the child’s estimate deviated from the actual target by 14.3%.

Mean PAE was calculated by taking the mean of all 18 trials (two trials of each of the nine target numbers). Children with lower mean PAE were considered to have more linear patterns of estimation than children with higher mean PAE. In this sample, $R^2_{lin}$ and mean PAE values were negatively correlated at the first ($r(17) = -.78, p < .001$) and second ($r(21) = -.86, p < .001$) time points, suggesting that PAE was a reasonable measure of linearity. Mean PAE was used to examine factors related to linearity of NLE performance (see Relationships amongst Variables below).

**Strategy Use.** Strategy use was inferred from patterns in the PAE in estimating individual target numbers, as Ashcraft and Moore (2012) and White and Szucs (2012) did. PAE for both trials of target numbers was calculated and graphed for individual children and average group performance, at both time points. Patterns were classified under one of five categories: tent-shaped, M-shaped, hook-shaped, low-error, or inconsistent. As per Ashcraft and Moore (2012), patterns of PAE were classified as tent-shaped if error was low at the anchors increasingly higher towards the midpoint (i.e., 5) and M-shaped if error was low at both anchors and the midpoint, with spikes of error in between. Additionally, patterns of PAE were classified as hook-shaped if error aligned with what Ashcraft and Moore (2012) observed in the 6-year-olds in their sample (but did not label or explore further). Patterns of PAE were classified as hook-shaped if error was low at the lower anchor and increased steadily thorough the third quartile (e.g., to 8), and (typically) declined near the upper anchor (e.g., at 9). Patterns of PAE were classified as low-error if error was uniformly below 20% (average between trials) for all target numbers. Patterns of PAE were classified as inconsistent if none of the other categories fit or if each trial could be classified under a different category (e.g., Trial 1 of target numbers showed an M-shape and Trial 2 showed a hook-shape). These categorizations were used to examine factors related to strategy use (see Relationships amongst Variables below). Group strategy use was also examined for group median PAE for each trial at each time point.

**Relationships amongst Variables.** The relationships between linearity of NLE performance and each of spatial skills, number knowledge and VMI and the relationship between these variables were examined using Pearson $r$ correlations for mean PAE, CMTT
The ability for spatial skills to predict growth in linearity of NLE performance was examined by correlating CMTT scores at Time 1 with absolute change in mean PAE from Time 1 to Time 2. Scatter plots were visually inspected for bivariate outliers (i.e., substantially off-trend), which were removed before correlations were calculated. The relationship between strategy use and each of spatial skills, number knowledge, VMI, and linearity of NLE performance were examined using a series of four one way ANOVAs with strategy categorization as the independent variable and CMTT (raw scores), NKT-A (raw scores), Beery VMI (raw scores) and mean PAE as dependent variables, respectively, at both time points. When the assumptions of parametric tests were violated, appropriate non-parametric tests were substituted.

**The Sample and Development.** A one sample t test was used to compare performance of the sample to normative data for the Beery VMI (standard score; Beery & Beery, 2010). One sample t tests were also used to compare performance of the sample to published data for children 4 years to 4 years 5 months at Time 1 and children 4 years 6 months to 4 years 11 months at Time 2 (Levine et al., 1999). Because we used only the 16 rotation items and Levine et al. (1999) used all 32 items (rotation and translation items), reported means were halved. This is a conservative threshold because the translation items would likely serve to inflate scores because they are easier than the rotation items. A chi square goodness-of-fit tests were used to compare the percentage of the sample who reached each developmental level on the NKT-A (i.e., AE score) to published percentages for 4-year-olds in Ontario from the NLSCY (Cycle 4; Lefebvre, Merrigan, & Roy-Desrosiers, 2011) at Time 1 and Time 2. Dependent samples t-tests were used to determine if significant gains in development had been made on the average Beery VMI (raw scores), CMTT, NKT-A (raw scores), and mean PAE.

**Results**

Group level descriptive statistics for each variable are presented in Table 2. The mean percent absolute difference in estimation (PADE) of target number between trials at each time point were calculated to determine if children’s estimations were consistent. Mean PADE below 20% was considered consistent. Most children’s target number estimations
between trials were consistent at each time point. At Time 1, the mean PADE of target numbers was over 20% for 3 of 20 children, with the group mean being 12.42% (SD = 7.49%). At Time 2, the mean PADE was over 20% for only 1 of 21 children, with the group mean being 10.00% (SD = 5.44%). Thus, children were more consistent in their estimates at the second time point.

**Model Selection**

Scatters of target number estimates plotted against target numbers (i.e., NLE patterns) were examined for bivariate outliers. At Time 1, one child (C20) had outlying values; at Time 2, four children (C5, C8, C12, C14) had outlying values. Figure 8 shows the NLE patterns for individual children at each time point with outliers marked. Outliers were removed before calculating $R^2$ values and AICc values to determine models of best fit for individual children. Tables 3 and 4 show $R^2$ and AICc values as well as the selected model for each child. The majority of children at Time 1 (76%) and Time 2 (67%) produced NLE patterns that were best fit by an unbounded power function. The majority of children (82%) had the same fit at both time points. Only two children showed what would be considered a developmental regression; that is, one child moved from a one-cycle power fit to an unbounded power fit (C1) and one child moved from a linear fit to an unbounded power fit. In line with this, when the NLE patterns for median target number estimates of the group at each time point were examined, an unbounded power function was the best fit (see Figure 9a).

**Linearity of Estimation**

$\beta$ values were only calculated for estimation patterns best fit by the unbounded power function because there was little representation for the other functions at either time point. These values as well as the direction of the change in these values can be seen in Table 5. It can be seen in Table 5 that half of those with estimation patterns best fit by the unbounded power function at both time points showed a decrease in $\beta$ towards greater linearity (i.e., approaching 1) and half showed an increase in $\beta$. The $\beta$ values for the unbounded power functions fitting the group median NLE patterns at Time 1 and Time 2 decreased from 1.56 to 1.39. The target estimates for the group became more linear over time.
Mean PAE was calculated for individual children at each time point as a measure of linearity (i.e., deviation from a perfect linear estimation; see Tables 3 and 4). Mean PAE decreased from Time 1 to Time 2 for 75% of children who completed the task at both time points (all but C1, C2, C14, C15, C21).

**Strategy Use**

PAE in estimating each target number on each trial was graphed for individual children at each time point. These graphs were visually examined for evidence of fitting one of four categories consistently: hook-shaped, tent-shaped, M-shaped, or low error. Graphs not fitting into any of these categories or fitting into a different one on each trial were placed in the inconsistent category. The categorization of the error patterns of individual children can be seen in Tables 3 and 4. Appropriateness of categorization choices were checked by graphing the mean PAE for each target number for the estimation data of the individuals placed into each of the five categories. These graphs can be seen in Figure 10. The hook-, tent- and M-shaped patterns are clearly visible, suggesting that the categorization of individual error patterns was appropriate. The majority of children either produced error patterns fitting the same category at each time point (45%) or produced an error pattern at Time 2 that showed more sophisticated strategy use than at Time 1 (e.g., hook-shaped to M-shaped; 35%; \( N = 20 \)). Two of the four children who showed “regression” in strategy use shifted to the inconsistent category; they both showed fairly low-error, but no discernible pattern.

Median PAE in estimating each target number on each trial was graphed for the group at each time point (see Figure 9b). At Time 1, the shape of the error pattern was fairly consistent between trials with the exception of the midpoint; Trial 1 would be categorized as M-shaped and Trial 2 would be categorized as tent-shaped. It can be inferred from this that on average children made use of the upper and lower anchors as reference points on both trials, but made use of the midpoint as an imagined reference point on the first trial only. This is consistent with the fact that only one child showed a consistent M-shaped error pattern on both trials at Time 1. The median PAE for Time 2 is nearly identical to Time 1 for target numbers 1, 2, 8, and 9 on both trials, but the shape of the group error pattern at Time 2 is slightly more difficult to categorize. Trial 2 would be categorized as low-error as median
PAE plateaus below 20%. Trial 1 would be categorized as tent-shaped because the greatest error was at the midpoint; however, there is clearly an M-shaped pattern, only with the dip in the M occurring at target number 7 rather than 5. This clear dip at 7 can also be seen in the inconsistent group and hook-shaped group graphs at Time 2 in Figure 10. It appears that children made use of the upper and lower anchor points on both trials. On Trial 1 it appears that 7, rather than 5, was used as an imagined reference point. One or more (e.g., 5 and 7) imagined reference points may have been used to achieve the low-error seen on Trial 2.

To examine how model of fit and strategy use align, the percentage of the children with NLE patterns best fit by an unbounded power model at Time 1 or Time 2 (N = 27) who fell into each strategy category was calculated. A hook-shaped error pattern (30%) and a tent-shaped error pattern (30%) aligned with an unbounded power fit most often, followed by inconsistent (15%), low-error (15%), and M-shaped (11%). The other fits also aligned with a variety of error patterns: one-cycle power (N = 4; 50% M-shaped, 50% low-error); logarithmic (N = 3; 67% inconsistent, 33% M-shaped); linear (N = 4; 50% inconsistent, 25% low-error, 25% hook); guessing/no fit (N = 3; 100% inconsistent). At the group level, NLE estimation patterns were best fit by unbounded power functions at both time points, but showed tent-shaped, M-shaped, and low-error patterns of median PAE. What was consistent across time points at the group level (and generally across individual children) was low error closer to the upper and lower anchors and greater error on the middle numbers.

**Relationships amongst Variables**

A series of Pearson $r$ correlations were used to examine the relationships between linearity of NLE performance, spatial skills, number knowledge, and VMI. Table 6 shows a correlation matrix displaying the relationship between mean PAE, CMTT (raw scores), NKT-A (raw scores), and Beery VMI (raw scores) across both time points. Bivariate outliers were removed based on visual inspection of scatter plots for points that were substantially off-trend.

Mean PAE at Time 1 significantly, negatively correlated to NKT-A scores at Time 1. Mean PAE at Time 2 was not related to any variables at either time point, including mean PAE at Time 1. One interpretation is that mean PAE is not a reliable measure of NLE
performance or that the NLE task is not reliable itself. However, most children did show
decreases in mean PAE over time, and for some the decrease was quite large (six children
showed decreases over 10% with one showing a decrease over 20%). When the four children
who unexpectedly showed an increase in PAE were removed from the analysis, the
correlation between mean PAE at Times 1 and 2 became significant, \( r(14) = .59, p = .015. \)
One of the children who showed an increase in mean PAE as well as another child who
showed virtually no decrease appeared to confuse the numerals ‘6’ and ‘9’ although the
numerals were read aloud to them (see C8 and C14 T2 in Figure 8). \( R^2 \text{lin} \) at Time 2 was also
not significantly related to NKT-A at Time 2, \( r(18) = .41, p = .07. \)

NKT-A at Time 1 was not related to any other variables at Time 1 except mean PAE,
though NKT-A scores at Time 1 were significantly, positively correlated with each of NKT-
A scores, CMTT scores, and Beery VMI scores at Time 2. NKT-A at Time 2 was not related
to any other variables except NKT-A at Time 1.

Beery VMI scores were significantly, positively correlated with NKT-A and CMTT
scores at Time 2. A similar relationship was observed at Time 1, but significance was not
reached. Notably, when the highest (C1) and lowest (C18) scores on the NKT-A and the
Beery VMI at both time points are treated as outliers and removed, the correlation between
Beery VMI scores and NKT-A scores at Time 2 is no longer significant, \( r(17) = .23, p = .34 \)
and the correlations between Beery VMI at Time 1 and NKT-A at Time 1 \( (r(16) = .06, p = .81) \) and Time 2 \( (r(17) = -.19, p = .44) \) became even weaker (see Figure 11).

CMTT scores were not related to NKT-A scores or mean PAE at either time points.
CMTT scores at Time 1 were not related to NKT-A scores or mean PAE at Time 2. In line
with this, CMTT scores at Time 1 did not predict absolute change in mean PAE from Time 1
to Time 2, \( r(19) = .16, p = .50. \)

A series of one-way ANOVAs were conducted to evaluate the relationship between
assigned strategy use category and spatial skills, numerical knowledge, VMI, and linearity of
NLE performance at both time points. Table 7 shows that only mean PAE on the NLE task
differed amongst the strategy groups at both time points. Tukey tests were used to evaluate
pairwise differences amongst the means, as the variances among the categories were similar.
At Time 1, the mean PAE of children with NLE error patterns classified as hook-shaped was higher than those with NLE error patterns classified as consistently low. At Time 2, the mean PAE of children with NLE error patterns classified as hook-shaped was higher than those with NLE error patterns classified as tent-shaped and those with NLE error patterns classified as consistently low.

**The Sample and Development**

A one-sample *t* test was conducted on Beery VMI standard scores at Time 1 to evaluate whether children in this sample differed significantly on VMI from children in the normative sample for the Beery VMI (*M* = 100). The sample mean of 110.57 (*SD* = 11.40) was significantly different from 100, *t*(20) = 4.25, *p* < .001. The 95% confidence interval for the mean difference in standard Beery VMI scores ranged from 5.38 to 15.76. The effect size, *d*, of .93 indicates a large effect. Children in the sample perform significantly higher on VMI than their same age peers. Children in the sample maintained this advantage at Time 2; average standard scores on the Beery VMI were nearly identical at Time 1 (*M* = 110.57, *SD* = 11.40) and Time 2 (*M* = 110.24, *SD* = 9.82), *t*(20) = -.16, *p* = .87.

One-sample *t* tests were conducted on CMTT scores at Time 1 to evaluate whether children in this sample differed significantly on CMTT from published averages for similar aged children (see Levine et al., 1999). At Time 1, the sample mean of 5.29 (*SD* = 2.74) was not significantly different from 4.98, *t*(20) = 0.51, *p* = .62. The 95% confidence interval for the mean difference in scores ranged from -.94 to 1.55. Using a conservative threshold, children in the current sample did not perform significantly better than children in the sample of Levine et al. (1999) at Time 1. At Time 2, the sample mean of 8.43 (*SD* = 1.99) was significantly different from 6.25, *t*(20) = 5.02, *p* < .001. The 95% confidence interval for the mean difference in scores ranged from 1.27 to 3.08. The effect size, *d*, of 1.10 indicates a large effect. Using a conservative threshold, children in the current sample performed significantly better than children in the sample of Levine et al. (1999) at Time 2.

Levine et al. (1999) reported a mean score of 9.96 out of 32 (all translation and rotation items) for children 4 years to 4 years 5 months. This score was halved for comparison since children in the current sample only received the 16 rotation items. This score was used for Time 1 when children in the current sample were 4 years 3 months on average. Likewise, Levine et al. (1999) reported a mean score of 12.50 out of 32 (all translation and rotation items) for children 4 years 6 months to 4 years 11 months. This score was halved and used for comparison at Time 2 when children in the current sample were 4 years 8 months on average.
A chi-square test of goodness-of-fit was performed to determine whether children in this sample (with a mean age of 4 years at both time points) differed significantly on NKT-AE scores at Time 1 and 2 from 4-year-olds children in Ontario (NLSCY Cycle 4; see Lefebvre, Merrigan, & Roy-Desrosiers, 2011). Lefebvre et al. (2011) reported that 47.98% of Ontario 4-year-olds passed the pre-dimensional level and 46.42% passed the uni-dimensional level. The percentage of children in this sample passing each level was not significantly different from the Ontario sample at Time 1 (65% and 35%, respectively), $\chi^2(1, N = 20) = 1.61, p = .21$, or Time 2 (38% and 57%, respectively), $\chi^2(1, N = 20) = .94, p = .33$.

A series of dependent samples $t$-tests were used to determine if significant gains in development had been made on average in the domains of spatial skills, number knowledge, VMI, and linearity of NLE. The results indicated that significant improvements occurred in each of these areas. Average scores on the CMTT were significantly higher at Time 2 ($M = 8.43, SD = 1.99$) than Time 1 ($M = 5.29, SD = 2.74$), $t(20) = 5.59, p < .001$. The standardized effect size index, $d$, was 1.22, which indicates a large effect. The 95% confidence interval for the mean difference between scores on the CMTT at Time 1 and 2 was 1.97 to 4.32. Average scores on the NKT-A were significantly higher at Time 2 ($M = 11.05, SD = 2.48$) than Time 1 ($M = 8.65, SD = 2.37$), $t(19) = 5.72, p < .001$. The standardized effect size index, $d$, was 1.28, which indicates a large effect. The 95% confidence interval for the mean difference between scores on the NKT-A at Time 1 and 2 was 1.52 to 3.28. Average raw scores on the Beery VMI were significantly higher at Time 2 ($M = 14.81, SD = 2.29$) than Time 1 ($M = 13.24, SD = 2.04$), $t(20) = 4.42, p < .001$. The standardized effect size index, $d$, was .96, which indicates a large effect. The 95% confidence interval for the mean difference between scores on the Beery VMI at Time 1 and 2 was .83 to 2.31. Mean PAE scores were significantly lower at Time 2 ($M = 13.20, SD = 5.35$) than Time 1 ($M = 18.01, SD = 8.68$), $t(19) = 5.72, p = .02$. The standardized effect size index, $d$, was .57, which indicates a medium effect. The 95% confidence interval for the mean decrease in mean PAE scores between Time 1 and 2 was .85 to 8.77. The correlations between scores at Time 1 and Time 2 for each of these variables, in Table 5, reveal that improvement was fairly consistent across children on the NKT-A and the Beery VMI, but less consistent across children on mean PAE and the CMTT.
Discussion

Does a strategy account better explain NLE performance than a logarithmic-to-linear shift account or a proportion-judgment account?

The results of this brief, longitudinal study did not support a logarithmic-to-linear shift account (e.g., Opfer et al., 2011). Few children’s individual NLE patterns were best fit by a logarithmic or linear function at either time point. Rather, the majority of children’s individual NLE patterns at both time points were best fit by an unbounded power function and the majority of children maintained the same fit from the beginning to the end of the study. Barth and Paladino (2011) found that the NLE patterns of 5-year-olds in their study could be fit equally well by logarithmic functions or power functions. That was not the case here. While some models were equally good fits statistically for some children, a logarithmic fit was a poor fit for most children. Rather than overestimating the magnitude of small numbers and underestimating the magnitude of larger numbers, as is indicative of the logarithmic fit, children in this sample tended to slightly underestimate both small and large numbers (and underestimate numbers around the midpoint to a great degree). A similar pattern was reported by Siegler and Opfer (2003) and Ashcraft and Moore (2012), but on the number-to-position version of the NLE task. This pattern was discussed by those authors as an exponential fit and considered equivalent to a logarithmic function (its inverse) because of the reversed nature of the task. Because a traditional position-to-number version of the NLE task was used in this study, there is no reason to consider the unbounded power fit observed to be equivalent to a logarithmic fit. Instead, the fit observed may be best explained by the strategies children used (discussed below) rather than being indicative of an internal representation of number that is linearly or logarithmically organized.

Despite the high proportion of children’s NLE patterns showing an unbounded power fit at both the beginning and the end of the study, the proportion-judgment account was also not supported for two main reasons. First, $\beta$ values of unbounded power functions were equally likely to shift towards or away from 1 (representing perfect linearity) from the beginning to the end of the study rather than approaching 1, as predicted by the proportion-judgment account (e.g., Barth & Paladino, 2011). It appears that $\beta$ values have similar limitations as $R^2_{lin}$, at least for unbounded power functions. A power function showing
increasing underestimation for all target numbers, but less underestimation at the upper anchor point (e.g., see C8 T1 in Figure 8) could have a $\beta$ value closer to 1 than a power function showing accurate estimation at both endpoints but inaccurate estimation approaching the midpoint (e.g., see C8 T2 in Figure 8). Comparing these $\beta$ values is like comparing the $R^2_{tin}$ of lines with different slopes; that is, $\beta$ values closer to 1 may not be related to greater accuracy. Mean PAE, used as the main measure of linearity of NLE in this study, overcomes this limitation. Unlike $\beta$ values, mean PAE values decreased for the majority of children between time points, suggesting development over time through maturation and experience.

Second, strategy use inferred from error patterns was not consistent with predictions made from the proportion-judgment account. Hollands and Dyre (2000) indicated that a one-cycle power model would result from using both anchor points to aid in estimation of target numbers. Most children in the current sample showed evidence of using both anchor points but only two children at each time point produced NLE patterns best fit by a one-cycle power model. Likewise, Hollands and Dyre (2000) indicated that when the midpoint is used as an additional reference point a two-cycle power model would be the best fit. In the current sample, several children showed evidence of using the midpoint to aid in estimation (i.e., an M-shaped error pattern) but the NLE patterns of no children were best fit by a two-cycle power model. It is possible that no two-cycle power fits were found here because the small range of target numbers estimated did not allow for four separate curves to be discerned.

There is at least one other possible reason for a failure to see one- and two-cycle power fits when evidence of the corresponding strategies existed. Hollands and Dyre (2000) noted that the one-cycle model could form an S in either direction (i.e., overestimation then underestimation, or vice versa); the cross through the midpoint is simply a product of a shift between over- and under-estimation, rather than accuracy at the midpoint due to strategic use of the midpoint. One question is: why must a shift occur? Could an unbounded power function result when overestimation (or underestimation) occurs at both end points? In this case, the curves of the S would fall on the same side and a cross at the midpoint would not occur because no shift between over- and under-estimation occurred. Slusser et al. (2013) pointed out that many factors could account for an unbounded power fit, but this was not a
possibility that was considered. Rather, Slusser et al. (2013) seemed to consider only factors that would account for unbounded power fits where there was inaccuracy at the upper anchor point, which was uncommon in the current sample. For example, Slusser et al. (2013) suggested that unbounded power fits could result from simply counting from left to right by randomly selected units.

While such counting behaviour was observed during the task, far more strategic counting behaviour was also observed. The examiner informally observed children in the current study using a number of counting strategies on the NLE task. Some children counted up from the lower anchor using their finger or a pencil without regard for unit size, as Slusser et al. (2013) suggested. Others were observed to use a counting strategy, but then make an adjustment at the end, seemingly through a visual check of plausibility. Some children were observed to count backwards from the upper anchor for larger numbers. Finally, at the second time point only, some children were observed to count forward from the lower anchor using their thumbs and forefingers to create consistent units, though the size of the unit was variable between target numbers. Thus, the high proportion of unbounded power fits seen could be due to a variety of strategic factors.

Indeed, since neither a logarithmic-to-linear shift nor a proportion-judgment account of NLE patterns were clearly supported, a strategy account was considered (White & Szucs, 2012). Clear evidence of strategy use was found by examining patterns of error for individual target numbers. Most children showed low error for numbers closer to the anchors, suggesting that these anchors were used to estimate the position of target numbers. Children producing hook-shaped error patterns seemed to use the upper anchor point in a limited way, sometimes using it to aid in the estimation of the target number 9. Children producing tent-shaped error patterns appeared to use the upper anchor point in a similar way to the lower anchor point; they used it to make increasingly more accurate estimations of target numbers 6-9. Some children also showed low percentages of error at the midpoint, suggesting use of an imagined reference point. Children producing M-shaped error patterns showed evidence of using one imagined reference point (the midpoint) and children producing a low-error error pattern may have used one or more imagined reference points to keep the error relatively low for all targets. On average, children in the inconsistent error
group showed little evidence of using the upper anchor or imagined reference points (though, some individual error patterns within this group showed evidence of one or both). Strategy use categories remained the same or shifted to a more sophisticated strategy at Time 2 (e.g., hook-shaped to M-shaped) for the majority of children.

Further support for the strategy account comes from the finding that linearity of NLE patterns differed amongst strategy groups. Those with a hooked-shaped pattern (denoting limited strategic use of the upper anchor) demonstrated significantly weaker linearity of NLE than those with tent-shaped patterns (denoting greater strategic use of the upper anchor) at Time 2 and those with low-error patterns (indicating strategic use of the upper anchor and possibly other imagined reference points) at both time points. One limitation in interpreting this finding is that the measure of linearity of NLE patterns used here was mean PAE and patterns of PAE for individual target numbers were used to determine strategy categories. By definition, children in the low-error category show generally low PAE. However, other measures of linearity (e.g., $\beta$, $R^2_{lin}$) have their own noted limitations.

At the group level, patterns of median error showed children using variable strategies between trials at both time points. At Time 1, children showed an M-shaped pattern on Trial 1 and a tent-shaped pattern on Trial 2. This suggests that strategy use became less sophisticated. An alternative explanation may lay in the administration of the NLE task. All children were given an orienting item and a training item on a 0-5 number line that was exactly half the size of the 0-10 number line used for the actual task. This may have alerted children to the location of the midpoint. This may have been enhanced by the fact that all children were given target numbers of 1 and 5 as their first two target numbers following the orienting and training items. By the second trial of the target number 5, the memory of the midpoint may have faded for some children, resulting in greater error at the midpoint. The training item may have also encouraged some children, who may not have done so independently, to use the upper anchor point. Children who placed the target number 4 to the left of the midpoint on the 0-5 number line were shown the orienting item again and told “It looks like the 4 should be a little closer to the 5.” This may have provided children with a strategy to use for the remainder of the task. While unintended, these observations lend more support to the strategy account. Children picked up on external cues and used them to
complete the task, rather than simply relying on an internal representation of number. Of course, children’s existing skills (spatial, numerical, VMI) could affect their ability to make use of these external cues.

It was predicted that if children were relying on an internal representation of number that their estimates of target numbers would be consistent from trial to trial. Most children’s target number estimations were consistent between trials at each time point. However, the remainder of the evidence presented thus far suggests that children were not relying on a consistent internal representation, but perhaps relying on strategies that yielded fairly consistent results.

**How are numerical knowledge, spatial skills, and NLE performance related?**

In line with predictions and past research showing a relationship between linearity of NLE and math abilities (Fischer et al., 2011; Booth & Siegler, 2008; Siegler & Booth, 2004; Siegler & Mu, 2008), linearity of NLE (i.e., mean PAE in estimation) was related to numerical knowledge at Time 1. However, linearity of NLE at Time 1 and 2 were not related to numerical knowledge at Time 2. One possible reason considered for failing to see an expected relationship at Time 2 was that the measure of linearity used here was not identical to the measure of linearity, $R_{Lin}^2$, used by others who found a relationship between linearity of NLE and other math abilities (e.g., Booth & Siegler, 2008; Siegler & Mu, 2008). Indeed, Booth and Siegler (2006) used mean PAE as a measure of accuracy alongside $R_{Lin}^2$ and found that $R_{Lin}^2$, but not mean PAE, was related to math achievement for children in Kindergarten through Grade 3. Geary (2011) also looked at the relationship between absolute error in NLE and math ability of children in Grade 1 and found that absolute error was not related to math ability in Grade 1 but did predict math ability in Grade 5. Thus, the literature suggests that mean PAE is less consistently related to math abilities than $R_{Lin}^2$. However, the difference in measure of linearity is likely not a key factor for two reasons: 1) mean PAE was related to numerical knowledge at Time 1; and 2) $R_{Lin}^2$ was also not related to numerical knowledge at Time 2.

A more likely reason for failing to see the expected relationship at Time 2 is that the range of numerical knowledge became more restricted. Indeed, most children (57%; n = 12)
earned one of only two scores (11, 12) at Time 2, while the same number of children earned scores ranging between 8 and 11 at Time 1. It is possible that effective, direct instruction in the requisite numerical knowledge (e.g., counting forward and backward, determining more and less, simple addition) brought most children up to the same level on the NKT-A. While children in the sample had a great deal of experience with numbers, they did not have any direct instruction on completing the number line task and thus, they may not have improved to meet a similar benchmark.

Contrary to predictions, there was no significant difference in the numerical knowledge of children demonstrating different strategies on the NLE task. Of course, this could mean that there truly is no relationship between strategy use on the NLE task and numerical knowledge. However, there are a number of other potential reasons for this finding. First, the NKT-A may not be sensitive to the kinds of numerical knowledge related to strategy use. The NKT-A measures some things that seem logically related to strategy use. For example, the ability to accurately determine more and less could help children establish which anchor a target number falls closest to and the ability to accurately count forward and backward could help children determine how far from the anchor the target number should be. As noted previously, such counting behaviour was observed. However, the NKT-A measures other skills, such as solving simple addition word problems, that may not be related to improved strategy use. A score representing a composite of these skills may not predict strategy use. Second, the five strategy categories may not have been distinct enough for differences in NKT-A to be seen. Differences between the five categories of strategy use mainly occur with respect to the middle numbers (see Figure 10). Most children successfully made use of the endpoints to some degree. It could be that the NKT-A did not tap into the kinds of skills (e.g., understanding proportions) that would promote the using the midpoint as an imagined reference point. At some level, the fact that children used quite similar strategies and had quite similar NKT-A scores suggests that numerical knowledge and strategy use may very well be observed to be related in a less homogenous sample and/or with a less familiar number range. The restricted range of variability in strategy use may have also account for a failure to see a relationship between strategy use and spatial ability or VMI.
Contrary to predictions, spatial ability was not related to numerical knowledge or linearity of NLE at either time point. Likewise, spatial ability at Time 1 was not predictive of numerical knowledge or linearity of NLE at Time 2, or changes in linearity of NLE from Time 1 to Time 2. Unexpectedly, numerical knowledge at Time 1 was related to spatial skills at Time 2.

One reason for spatial abilities at Time 1 not having predictive value may be that 28% of children performed at or below chance on the CMTT task at this time point with an additional 24% performing just above chance (i.e., 5/16). These low scores could reflect an inability to perform the spatial task. The low scores could also relate to the test-taking behaviour of young children. In testing situations, young children may be more likely to be uncomfortable, poorly motivated, or inattentive to the task (Ginsburg & Pappas, 2004). These qualities were obvious in the sample on whole in October, leading to the decision to postpone initial testing to December. The multiple-choice nature of this particular task may have further encouraged impulsive responding. With such a small sample size, inconsistent responding by just a few children could have a large impact on analyses. The low scores could also reflect the difficulty that young children may have understanding such an abstract task. Indeed, the instructions for this task are brief and no training items or feedback are used. Children may have better been able to understand the intention of the task at Time 2 because they explored symmetry in their classroom between the time points and the task involves putting symmetrical halves together to form a whole. The restricted range of NKT-A scores at Time 2 may have prevented a correlation between spatial and numerical scores at Time 2, resulting in only numerical knowledge at Time 1 and spatial skills at Time 2 being correlated.

Another reason why spatial abilities may not have been related to numerical knowledge or linearity of NLE in this sample is the sample was restricted to children from higher-SES families. In other studies where relationships between early spatial and numeracy skills were found mixed-SES samples were used (e.g., Barnes et al., 2011; Farmer et al., 2013; Gunderson et al., 2012). Farmer et al. (2013) found SES differences on spatial and numerical measures in 3 year old children that persisted through to age 5. Without lower-SES children in the current sample, the variability in scores may have been too restricted to see the same
kind of relationship between abilities. The variability was likely further restricted in the current sample due to the high-quality instruction all children received.

**Is VMI a confound in the NLE task or a correlate that increases its value in predicting other math abilities?**

Previous research has shown mixed evidence with regards to the relationship between VMI and math ability (Kurdek & Sinclair, 2001; Sortor & Kulp, 2003). In the current study, when scores for all children were included, VMI was significantly related to numerical knowledge at Time 2. A similar pattern, only approaching significance, was observed for VMI and numerical knowledge at Time 1 and VMI at Time 1 and numerical knowledge at Time 2. However, when the scores of the two children with the highest and lowest demonstrated numerical knowledge and VMI at both time points were removed from the analyses, no relationships were observed. This finding adds strength to arguments made thus far about the role having such a homogenous sample played in failure to observe expected significant relationships between variables. The trend observed here suggests that with a less homogenous sample containing additional lower-performing children, a relationship between VMI and numerical knowledge may have been observed. It is possible that the inconsistencies seen in limited literature on the relationship between VMI and math ability are due to the samples examined. Indeed, Kurdek and Sinclair (2001), who did not find a relationship between VMI and math ability, expressed concerns that the results from their primarily middle-class sample may not have generalized to more diverse samples.

To the best of my knowledge, this is the first study in which the relationship between VMI and NLE performance was examined. In this sample, VMI and NLE were not related. Figure 12, showing the NLE markings of two children for the same number, demonstrates the concern that led to the exploration of this relationship. In Figure 12a the child’s marking was nearly perfectly straight, whereas in Figure 12b, the difference between where the child began the mark and where it crossed the number line is 15 mm. Given that the space from one digit to the next is only 18.5 mm, this could result in a lot of error in estimation that may or may not be related to numerical skill. No conclusions can be drawn about the role of VMI in NLE performance of this sample.
Description of the Development of Children Learning in a Play- and Inquiry-Based Classroom

The children in the current sample were learning in a relatively unique environment: a laboratory school. The classroom was a child-centered, play- and inquiry-based environment where teaching practices in mathematics were aligned with the recommendations of national organizations. Specifically, the description of the classroom and specific examples provided here demonstrate at least six of the ten recommendations for educators proposed by the NAEYC and NCTM (2010) were met. The classroom teacher provided “ample time, materials, and teacher support for children to engage in play” (Recommendation 8) and built on “children’s natural interest in mathematics” (Recommendation 1, p. 3). She also set out to “actively introduce mathematical concepts, methods, and language” (Recommendation 9), “provide for children’s deep and sustained interaction with key mathematical ideas” (Recommendation 6), and “use teaching practices that strengthen children’s problem-solving and reasoning processes” (Recommendation 4, p. 3). Further, she integrated “mathematics with other activities and other activities with mathematics” (Recommendation 7, p. 3). At the school level, structures were in place to encourage continual learning and collaboration amongst teachers and other professionals. Such institutional support is crucial in helping teachers provide high quality math education to young children (NAEYC and NCTM, 2010).

There was some evidence that children learning in this environment had better than typical VMI and spatial skills (near the end of the school year only), but that their basic numerical knowledge was not significantly better than typical. Spatial skills and VMI may have been better developed in this sample of children because of the strong focus on purposeful, sustained play in their classroom, particularly in the areas of building and completing puzzles. Opportunities to develop these skills are likely not as readily available in the early learning environments of many children (Clements & Sarama, 2011). When math is addressed in early learning environments, basic numerical skills are often a focus (Clements & Sarama, 2011). Within Ontario (Ontario Ministry of Education, 2006), there is an explicit expectation that teachers will address the kinds of concepts that are evaluated at the pre-dimensional and uni-dimensional levels of the NKT-A, a measure on which the current sample was compared to an Ontario sample. While the children in this study had the
benefit of some rich numerical learning opportunities, the knowledge gained through these experiences were not addressed on the NKT-A.

On average, children in the sample made significant gains in all areas measured. The gains seen in spatial skills, numerical knowledge, and VMI were quite large (effect sizes, $d$, ranging from 0.96 to 1.28) and the gains seen in linearity of NLE were moderate ($d = .57$). Improvement was fairly consistent across children in their numerical knowledge and VMI but less consistent across children on mean PAE and spatial skills. As noted previously, the lack of consistency seen in performance on these tasks could be related to the young age of the children and the effect it has on their motivation, attention, impulsivity (Ginsburg & Pappas, 2004), and ability to understand the intention of such abstract tasks. The lack of consistency seen in growth on these tasks could also be related to a true lack of consistency in growth on the underlying skills measured. For example, while efforts were made to engage all children in the classroom in spatially-related play, some children naturally gravitated towards these activities during free-play and thus had greater opportunities to use and build their skills. Importantly, it must be noted that while an effort was made to compare the current sample to other available samples, without a true control group, it is difficult to draw firm conclusions about the amount of growth that was due to high-quality teaching and the amount of growth that was due to other factors, such as maturation and stimulation at home. That being said, the children made substantial gains in their skills over such a short time period and these gains appeared larger than might be seen in a more typical environment.

**Future Research and Conclusions**

The results of this study did not support a logarithmic-to-linear shift account of NLE performance. This study adds to a growing body of evidence (e.g., Barth & Paladino, 2011; Boumeester & Verkoeijen, 2013; Slusser et al., 2013; White & Szucs, 2012) suggesting that NLE performance should not be interpreted as directly indicative of internal representations of number, as it almost exclusively has been until recently. Based on the interpretation that the linearity of children’s internal representations of number is predictive of math achievement (e.g., Booth & Siegler, 2006; Siegler & Booth, 2004), math learning disabilities have been characterized as a developmentally inappropriate overreliance on a logarithmic
representation of number (Geary et al., 2007, 2008) and training studies have been developed with the intention of improving the linearity of representations (e.g., Fischer et al., 2011; Siegler & Ramani, 2009; Whyte & Bull, 2008). Thus, over-interpreting NLE patterns has real implications for clinical and educational settings.

Based on the results of this study, I agree with the conclusion of White and Szucs (2012) that it may not be useful to interpret a best fitting model for NLE patterns at all. Rather, examining strategy use, inferred from patterns of error, may be more informative. Clear evidence of strategy use was observed in this sample and the kinds of strategies employed were related to accuracy in estimation, which improved over time. Accuracy of estimation may provide a better indication of children’s understanding of the linear number line than other measures of linearity (e.g., $R^2_{lin}, \beta$). Unfortunately, only one of the predicted relationships between strategy use, accuracy in estimation, spatial skills, and numerical knowledge that would have provided further evidence for a strategy account of NLE was supported: accuracy in NLE and numerical knowledge were related at the beginning of the study. Likewise, the question of whether VMI is a confound the NLE task remains open for exploration. A major factor in the failure to see relationships was likely the homogeneity of the sample selected (mostly higher-SES and high-achieving).

The strategy account of NLE performance and the roles of spatial abilities and VMI in strategy use and accuracy in NLE should be investigated further. Selecting a less homogenous sample and using a larger, less familiar number range may provide the variability necessary to draw clearer conclusions. A finer grained analysis of strategy use may be possible if individual numerical tasks, as opposed to a composite measure, was used. It may be fruitful to examine understanding of the stable-order principle of counting, cardinality, numeral recognition, “more” and “less”, and halving and quartering within the selected range. Further, clinical interviewing could be useful in overcoming some of the difficulties implicit in using standardized measures with young children (Ginsburg & Pappas, 2004) and could allow for a richer understanding of children’s understanding of the linear number line and the strategies they employ (e.g., asking questions such as: “How did you know to put the mark there?”). There may be benefit in examining a variety of spatial abilities, beyond and including the intrinsic-dynamic spatial abilities examined here.
Specifically, it may be beneficial to employ the supplementary subtests of the Beery VMI that allows for discernment between the spatial (intrinsic-static) and fine-motor aspects of VMI. It could be that neither, only one, or the integration of the two are influential. Finally, both stability (use of endpoints) and growth (reduction in overall error) in NLE performance was observed over a five month period. Longer longitudinal studies will further inform our understanding of how knowledge of the linear number line develops.

While the conclusions that can be drawn about the effects of the unique learning environment on growth of this sample are limited, one important conclusion that can be drawn is that young children are ready for and interested in learning math concepts and engaging in meaningful problem solving using math. With careful attention and planning on the part of the teacher, an environment can be created where the line between math and play is blurred, or even non-existent. Furthermore, immense gains in mathematical understanding can be achieved in such an environment. Given how critical early math learning is to later academic achievement (e.g., Duncan et al., 2007), concrete examples of classrooms where this goal is achieved are an valuable resource for educators of young children, particularly for those who see intentional teaching of math concepts as contraindicated in a play-based program. In Ontario, where a new all day, play-based Kindergarten curriculum has recently been rolled out and thousands of teachers are reconsidering what their Kindergarten programs should look like, an example such as this is timely and important.
References


### Table 1

*Math Topics Intentionally Addressed between December and May in the Areas of Numerical and Spatial Skills*

<table>
<thead>
<tr>
<th>Numerical Skills</th>
<th>Spatial Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Counting (backwards and forwards)</td>
<td>• Sorting 2D and 3D shapes by common properties</td>
</tr>
<tr>
<td>• Skip counting (e.g., by 2s, by 5s)</td>
<td>• Noticing and creating symmetry (2D and 3D)</td>
</tr>
<tr>
<td>• Determining “more” and “less”</td>
<td>• Translation and rotating shapes (2D and 3D)</td>
</tr>
<tr>
<td>• Composing and decomposing numbers&lt;sup&gt;a&lt;/sup&gt;</td>
<td>• Composing and decomposing shapes (2D and 3D)</td>
</tr>
<tr>
<td>• Adding, subtracting, multiplying, and dividing within a narrative context&lt;sup&gt;b&lt;/sup&gt; using concrete materials (e.g., fingers)</td>
<td>• Coverting 2D representations to 3D representations, and vice versa</td>
</tr>
<tr>
<td>• Abstract computation (addition and subtraction)</td>
<td>• Patterning&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>• Standard and non-standard measurement&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* 2D = two-dimensional; 3D = three-dimensional.  
<sup>a</sup>For example, children may determine all of the ways you can make 7.  
<sup>b</sup>For example, they may be asked: “How many more blocks would you need to make the towers equal?”; “If there are two cows, how many legs would there be?”; or “How many people should be on each team to make them fair?”.

<sup>c</sup>This is technically a separate strand of math but measurement problems provide a context for children to apply their understanding of counting, more/less etc.  
<sup>d</sup>This is technically a separate strand of math but requires spatial skills, especially for more complex patterns.
Table 2

**Descriptive Statistics for All Variables**

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M (SD)</td>
</tr>
<tr>
<td>Beery VMI (raw)</td>
<td>21</td>
<td>13.42 (2.05)</td>
</tr>
<tr>
<td>Beery VMI (standard)</td>
<td>21</td>
<td>110.57 (11.40)</td>
</tr>
<tr>
<td>CMTT (raw)</td>
<td>21</td>
<td>5.29 (2.74)</td>
</tr>
<tr>
<td>NKT-A (raw)</td>
<td>20</td>
<td>8.65 (2.37)</td>
</tr>
<tr>
<td>Mean PAE</td>
<td>20</td>
<td>18.01 (8.68)</td>
</tr>
</tbody>
</table>
Table 3  
*Summary of Number Line Estimation Data for Individual Children at Time 1*

<table>
<thead>
<tr>
<th>Child</th>
<th>Log</th>
<th>Linear</th>
<th>Unbound</th>
<th>1-cycle</th>
<th>2-Cycle</th>
<th>Log</th>
<th>Linear</th>
<th>Unbound</th>
<th>1-cycle</th>
<th>2-Cycle</th>
<th>Model of Best Fit</th>
<th>Mean PAE</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
<td>0.67</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>33.11</td>
<td>30.40</td>
<td>29.90</td>
<td>27.52</td>
<td>27.08</td>
<td>1-cycle</td>
<td>15.26</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.95</td>
<td>0.95</td>
<td>0.91</td>
<td>0.90</td>
<td>4.56</td>
<td>-13.71</td>
<td>-15.20</td>
<td>-6.40</td>
<td>-6.27</td>
<td>Linear</td>
<td>6.55</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>22.49</td>
<td>Incon</td>
</tr>
<tr>
<td>4</td>
<td>0.54</td>
<td>0.72</td>
<td>0.86</td>
<td>0.16</td>
<td>0.13</td>
<td>23.89</td>
<td>15.42</td>
<td>2.17</td>
<td>32.36</td>
<td>33.00</td>
<td>Unbound</td>
<td>19.76</td>
<td>Tent</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.86</td>
<td>0.95</td>
<td>0.72</td>
<td>0.72</td>
<td>20.27</td>
<td>5.88</td>
<td>12.18</td>
<td>15.87</td>
<td>15.67</td>
<td>Unbound</td>
<td>11.74</td>
<td>Tent</td>
</tr>
<tr>
<td>6</td>
<td>0.61</td>
<td>0.74</td>
<td>0.81</td>
<td>0.79</td>
<td>x</td>
<td>33.42</td>
<td>26.31</td>
<td>20.55</td>
<td>19.31</td>
<td>x</td>
<td>Unbound</td>
<td>16.19</td>
<td>Tent</td>
</tr>
<tr>
<td>7</td>
<td>0.85</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>7.48</td>
<td>-6.80</td>
<td>-6.82</td>
<td>-9.77</td>
<td>-11.48</td>
<td>1-cycle</td>
<td>5.35</td>
<td>Low</td>
</tr>
<tr>
<td>8</td>
<td>0.48</td>
<td>0.58</td>
<td>0.61</td>
<td>x</td>
<td>x</td>
<td>3.48</td>
<td>-0.12</td>
<td>-1.76</td>
<td>47.84</td>
<td>x</td>
<td>Unbound</td>
<td>34.44</td>
<td>Hook</td>
</tr>
<tr>
<td>9</td>
<td>0.66</td>
<td>0.55</td>
<td>0.60</td>
<td>0.54</td>
<td>x</td>
<td>19.01</td>
<td>24.21</td>
<td>22.02</td>
<td>x</td>
<td>22.14</td>
<td>Log</td>
<td>13.75</td>
<td>Incon</td>
</tr>
<tr>
<td>10</td>
<td>0.69</td>
<td>0.80</td>
<td>0.80</td>
<td>-0.06</td>
<td>x</td>
<td>6.92</td>
<td>-0.97</td>
<td>-1.18</td>
<td>26.64</td>
<td>x</td>
<td>Unbound</td>
<td>17.69</td>
<td>Hook</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>33.63</td>
<td>Incon</td>
</tr>
<tr>
<td>12</td>
<td>0.77</td>
<td>0.87</td>
<td>0.88</td>
<td>0.52</td>
<td>0.49</td>
<td>8.48</td>
<td>-2.26</td>
<td>-4.21</td>
<td>18.73</td>
<td>19.91</td>
<td>Unbound</td>
<td>13.99</td>
<td>Incon</td>
</tr>
<tr>
<td>14</td>
<td>0.71</td>
<td>0.83</td>
<td>0.86</td>
<td>0.83</td>
<td>0.83</td>
<td>23.38</td>
<td>13.67</td>
<td>10.63</td>
<td>10.60</td>
<td>10.92</td>
<td>Unbound</td>
<td>10.48</td>
<td>Incon</td>
</tr>
<tr>
<td>15</td>
<td>0.52</td>
<td>0.65</td>
<td>0.69</td>
<td>0.25</td>
<td>0.17</td>
<td>23.88</td>
<td>17.88</td>
<td>15.80</td>
<td>29.12</td>
<td>31.00</td>
<td>Unbound</td>
<td>16.67</td>
<td>Hook</td>
</tr>
<tr>
<td>16</td>
<td>0.65</td>
<td>0.83</td>
<td>0.96</td>
<td>0.79</td>
<td>0.77</td>
<td>27.18</td>
<td>14.04</td>
<td>-13.81</td>
<td>15.34</td>
<td>16.63</td>
<td>Unbound</td>
<td>12.19</td>
<td>Tent</td>
</tr>
<tr>
<td>17</td>
<td>0.40</td>
<td>0.54</td>
<td>0.80</td>
<td>-0.26</td>
<td>-0.31</td>
<td>30.22</td>
<td>25.15</td>
<td>10.68</td>
<td>40.86</td>
<td>41.62</td>
<td>Unbound</td>
<td>25.32</td>
<td>Hook</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>34.35</td>
<td>Incon</td>
</tr>
<tr>
<td>19</td>
<td>0.63</td>
<td>0.80</td>
<td>0.93</td>
<td>0.50</td>
<td>0.50</td>
<td>22.72</td>
<td>11.95</td>
<td>-6.45</td>
<td>25.39</td>
<td>25.51</td>
<td>Unbound</td>
<td>16.67</td>
<td>Incon</td>
</tr>
<tr>
<td>20</td>
<td>0.52</td>
<td>0.68</td>
<td>0.77</td>
<td>-0.58</td>
<td>-0.84</td>
<td>16.89</td>
<td>10.05</td>
<td>4.12</td>
<td>34.48</td>
<td>37.12</td>
<td>Unbound</td>
<td>23.96</td>
<td>Hook</td>
</tr>
<tr>
<td>21</td>
<td>0.73</td>
<td>0.85</td>
<td>0.89</td>
<td>0.83</td>
<td>0.84</td>
<td>20.88</td>
<td>9.48</td>
<td>3.73</td>
<td>9.25</td>
<td>8.88</td>
<td>Unbound</td>
<td>9.76</td>
<td>Low</td>
</tr>
</tbody>
</table>

Note. AICc = Akaike Information Criteria Corrected for small sample sizes; PAE = Percent Absolute Error in Estimation; - = Appeared to be guessing, so no patterns of fit examined; x = incalculable or illogical value returned. *One child (C13) did not complete this task. Five models of best fit were considered as the best fit for target number estimates plotted against target numbers: logarithmic (log); linear; unbounded power (unbound); one-cycle power (1-cycle); two-cycle power (2-cycle). Best fit was based on the lowest AICc value unless statistically equivalent (differed by less than two points), then visual inspection was used for selection. Strategy was inferred from patterns of PAE for individual target numbers across trials. Five strategy groups were considered: hook-shaped (Hook); tent-shaped (Tent); M-shaped (M), low-error (Low); and inconsistent error (Incon).
Table 4

Summary of Number Line Estimation Data for Individual Children at Time 2

<table>
<thead>
<tr>
<th>Child</th>
<th>Log</th>
<th>Linear</th>
<th>Unbound</th>
<th>1-cycle</th>
<th>2-Cycle</th>
<th>Log</th>
<th>Linear</th>
<th>Unbound</th>
<th>1-cycle</th>
<th>2-Cycle</th>
<th>Model of Best Fit</th>
<th>Mean PAE</th>
<th>Strategya</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45</td>
<td>0.59</td>
<td>0.76</td>
<td>-0.12</td>
<td>-0.18</td>
<td>27.93</td>
<td>22.53</td>
<td>13.27</td>
<td>38.18</td>
<td>39.11</td>
<td>Unbound</td>
<td>23.54</td>
<td>Hook</td>
</tr>
<tr>
<td>2</td>
<td>0.76</td>
<td>0.91</td>
<td>0.96</td>
<td>0.83</td>
<td>0.84</td>
<td>15.56</td>
<td>-1.94</td>
<td>-15.12</td>
<td>6.24</td>
<td>5.69</td>
<td>Unbound</td>
<td>9.28</td>
<td>Low</td>
</tr>
<tr>
<td>3</td>
<td>0.71</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>0.69</td>
<td>11.84</td>
<td>7.39</td>
<td>7.92</td>
<td>4.75</td>
<td>10.62</td>
<td>1-Cycle</td>
<td>9.58</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>0.61</td>
<td>0.79</td>
<td>0.92</td>
<td>0.41</td>
<td>0.40</td>
<td>21.23</td>
<td>9.58</td>
<td>-8.29</td>
<td>25.89</td>
<td>26.28</td>
<td>Unbound</td>
<td>16.46</td>
<td>Tent</td>
</tr>
<tr>
<td>5</td>
<td>0.73</td>
<td>0.88</td>
<td>0.97</td>
<td>0.87</td>
<td>0.86</td>
<td>21.19</td>
<td>6.76</td>
<td>-14.12</td>
<td>6.27</td>
<td>7.56</td>
<td>Unbound</td>
<td>11.32</td>
<td>Tent</td>
</tr>
<tr>
<td>6</td>
<td>0.78</td>
<td>0.92</td>
<td>0.96</td>
<td>0.83</td>
<td>0.83</td>
<td>12.20</td>
<td>-6.55</td>
<td>-18.82</td>
<td>5.57</td>
<td>5.36</td>
<td>Unbound</td>
<td>9.49</td>
<td>Low</td>
</tr>
<tr>
<td>7</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
<td>-11.62</td>
<td>-13.09</td>
<td>-16.19</td>
<td>-15.44</td>
<td>-13.18</td>
<td>1-Cycle</td>
<td>5.32</td>
<td>Low</td>
</tr>
<tr>
<td>8</td>
<td>0.57</td>
<td>0.70</td>
<td>0.76</td>
<td>-0.68</td>
<td>x</td>
<td>14.58</td>
<td>8.82</td>
<td>5.12</td>
<td>33.61</td>
<td>x</td>
<td>Unbound</td>
<td>25.20</td>
<td>Hook</td>
</tr>
<tr>
<td>9</td>
<td>0.72</td>
<td>0.59</td>
<td>0.66</td>
<td>0.55</td>
<td>0.46</td>
<td>12.10</td>
<td>19.20</td>
<td>15.94</td>
<td>18.14</td>
<td>21.52</td>
<td>Log</td>
<td>12.91</td>
<td>Incon</td>
</tr>
<tr>
<td>10</td>
<td>0.77</td>
<td>0.85</td>
<td>0.86</td>
<td>0.74</td>
<td>0.74</td>
<td>11.95</td>
<td>4.28</td>
<td>3.12</td>
<td>11.93</td>
<td>11.87</td>
<td>Unbound</td>
<td>10.39</td>
<td>Incon</td>
</tr>
<tr>
<td>11</td>
<td>0.54</td>
<td>0.58</td>
<td>0.58</td>
<td>0.54</td>
<td>0.50</td>
<td>22.10</td>
<td>20.26</td>
<td>20.27</td>
<td>19.27</td>
<td>20.95</td>
<td>Linear</td>
<td>14.77</td>
<td>Incon</td>
</tr>
<tr>
<td>12</td>
<td>0.72</td>
<td>0.87</td>
<td>0.93</td>
<td>0.68</td>
<td>0.68</td>
<td>16.53</td>
<td>3.69</td>
<td>-6.51</td>
<td>16.13</td>
<td>16.03</td>
<td>Unbound</td>
<td>13.57</td>
<td>Tent</td>
</tr>
<tr>
<td>13</td>
<td>0.60</td>
<td>0.50</td>
<td>0.55</td>
<td>0.47</td>
<td>0.36</td>
<td>20.19</td>
<td>23.75</td>
<td>21.96</td>
<td>22.38</td>
<td>25.54</td>
<td>Log</td>
<td>19.88</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>0.63</td>
<td>0.79</td>
<td>0.96</td>
<td>0.60</td>
<td>0.60</td>
<td>25.03</td>
<td>15.69</td>
<td>-10.78</td>
<td>23.56</td>
<td>23.57</td>
<td>Unbound</td>
<td>17.00</td>
<td>M</td>
</tr>
<tr>
<td>15</td>
<td>0.46</td>
<td>0.60</td>
<td>0.77</td>
<td>0.02</td>
<td>-0.01</td>
<td>29.46</td>
<td>23.75</td>
<td>13.57</td>
<td>37.47</td>
<td>38.06</td>
<td>Unbound</td>
<td>22.31</td>
<td>Hook</td>
</tr>
<tr>
<td>16</td>
<td>0.77</td>
<td>0.88</td>
<td>0.91</td>
<td>0.87</td>
<td>0.87</td>
<td>16.89</td>
<td>4.20</td>
<td>0.13</td>
<td>3.40</td>
<td>3.87</td>
<td>Unbound</td>
<td>8.20</td>
<td>Tent</td>
</tr>
<tr>
<td>17</td>
<td>0.80</td>
<td>0.90</td>
<td>0.92</td>
<td>0.89</td>
<td>0.89</td>
<td>15.37</td>
<td>3.57</td>
<td>-1.43</td>
<td>1.41</td>
<td>2.87</td>
<td>Unbound</td>
<td>8.20</td>
<td>M</td>
</tr>
<tr>
<td>18</td>
<td>0.66</td>
<td>0.68</td>
<td>0.68</td>
<td>0.64</td>
<td>0.60</td>
<td>18.54</td>
<td>17.57</td>
<td>17.46</td>
<td>17.38</td>
<td>18.90</td>
<td>Linear</td>
<td>13.63</td>
<td>Hook</td>
</tr>
<tr>
<td>19</td>
<td>0.72</td>
<td>0.88</td>
<td>0.98</td>
<td>0.84</td>
<td>0.83</td>
<td>22.28</td>
<td>7.64</td>
<td>-22.75</td>
<td>9.29</td>
<td>11.18</td>
<td>Unbound</td>
<td>11.47</td>
<td>Low</td>
</tr>
<tr>
<td>20</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
<td>0.67</td>
<td>0.62</td>
<td>6.59</td>
<td>6.74</td>
<td>6.13</td>
<td>11.79</td>
<td>14.59</td>
<td>Linear</td>
<td>10.18</td>
<td>Incon</td>
</tr>
<tr>
<td>21</td>
<td>0.68</td>
<td>0.83</td>
<td>0.90</td>
<td>0.78</td>
<td>0.77</td>
<td>23.97</td>
<td>12.48</td>
<td>3.00</td>
<td>14.77</td>
<td>15.26</td>
<td>Unbound</td>
<td>11.14</td>
<td>M</td>
</tr>
</tbody>
</table>

Note. AICc = Akaike Information Criteria Corrected for small sample sizes; PAE = Percent Absolute Error in Estimation; x = incalculable or illogical value returned. *Five models of best fit were considered as the best fit for target number estimates plotted against target numbers: logarithmic (log); linear; unbounded power (unbound); one-cycle power (1-cycle); two-cycle power (2-cycle). Best fit was based on the lowest AICc value unless statistically equivalent (differed by less than two points), then visual inspection was used for selection. *Strategy was inferred from patterns of PAE for individual target numbers across trials. Five strategy groups were considered: hook-shaped (hook); tent-shaped (tent); M-shaped (M), low-error (low); and inconsistent error (incon).
Table 5

\(\beta\) Values for Number Line Estimation Patterns of Children that were Best Fit by an Unbounded Power Function at each Time Point and the Direction of the Change

<table>
<thead>
<tr>
<th>Child</th>
<th>T1</th>
<th>T2</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2.88</td>
<td>n/a</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1.43</td>
<td>n/a</td>
</tr>
<tr>
<td>4</td>
<td>2.25</td>
<td>2.05</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>1.70</td>
<td>1.66</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1.61</td>
<td>1.37</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1.49</td>
<td>1.71</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>1.08</td>
<td>1.17</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>1.20</td>
<td>1.55</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>1.31</td>
<td>2.25</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>1.51</td>
<td>2.71</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>1.94</td>
<td>1.28</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>4.45</td>
<td>1.28</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>2.02</td>
<td>1.72</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>2.00</td>
<td></td>
<td>n/a</td>
</tr>
<tr>
<td>21</td>
<td>1.40</td>
<td>1.58</td>
<td>+</td>
</tr>
</tbody>
</table>

*Note.* n/a = not applicable; - = decrease in \(\beta\) values; + = increase in \(\beta\) values
Table 6  
**Correlations between Variables at Time 1 and Time 2**

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean PAE  NKT-A  CMTT Beery VMI</td>
<td>Mean PAE  NKT-A  CMTT Beery VMI</td>
</tr>
<tr>
<td>Mean PAE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NKT-A</td>
<td>-.67**(p = .001)  .22 (p = .352)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(n = 20)</td>
<td>(n = 20)</td>
</tr>
<tr>
<td>CMTT</td>
<td>-.19 (p = .434)  .22 (p = .352)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(n = 20)</td>
<td>(n = 20)</td>
</tr>
<tr>
<td>Beery VMI</td>
<td>-.34 (p = .145)  .41* (p = .072)  .42 (p = .056)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(n = 20)</td>
<td>(n = 20)</td>
</tr>
</tbody>
</table>

**Time 2**

|                  |                      |                      |
| Mean PAE         | .35 \(p = .133\)  -.23 \(p = .337\)  .02 \(p = .945\)  -.35 \(p = .132\)  |                      |
|                  | (n = 20)             | (n = 19)             | (n = 21)          | (n = 20)          |
| NKT-A            | -.45 \(p = .053\)  .70** \(p = .001\)  .23 \(p = .330\)  .42* \(p = .056\)  -.29 \(p = .220\)  |                      |
|                  | (n = 19)             | (n = 20)             | (n = 20)          | (n = 21)          | (n = 20)          |
| CMTT             | -.41 \(p = .076\)  .68** \(p = .001\)  .44* \(p = .044\)  .34 \(p = .129\)  -.14 \(p = .549\)  .29 \(p = .223\)  |                      |
|                  | (n = 20)             | (n = 19)             | (n = 21)          | (n = 21)          | (n = 20)          | (n = 20)          |
| Beery VMI        | -.31 \(p = .179\)  .51* \(p = .021\)  .42 \(p = .062\)  .72** \(p < .001\)  -.45 \(p = .053\)  .58**a \(p = .006\)  .61** \(p = .003\)  |                      |
|                  | (n = 20)             | (n = 20)             | (n = 21)          | (n = 21)          | (n = 19)          | (n = 21)          | (n = 21)          |

*Note.* *When the one child with the highest (C1) and lowest (C18) NKT-A and Beery VMI scores at both time points are removed from the analysis, \(p\) values greatly increase, such that none of the three correlations are significant.  *\(p < .05; **p < .01.*
Table 7  
*Comparisons of Strategy Groups on Spatial, Numerical, and VMI Skills at Time 1 and Time 2*

<table>
<thead>
<tr>
<th>Strategy Groups&lt;sup&gt;a&lt;/sup&gt;</th>
<th>M (SD)</th>
<th>F</th>
<th>(\eta^2)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incon</td>
<td>Hook</td>
<td>Tent</td>
<td>M</td>
</tr>
<tr>
<td>Time 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df = 3</td>
<td>N = 7</td>
<td>N = 5</td>
<td>N = 4</td>
<td>N = 1&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean PAE</td>
<td>20.77 (9.75)</td>
<td>23.61 (7.13)</td>
<td>14.97 (3.77)</td>
<td>15.26</td>
</tr>
<tr>
<td>NKT-A</td>
<td>8.00 (3.06)</td>
<td>7.60 (.89)</td>
<td>9.25 (1.26)</td>
<td>13.00</td>
</tr>
<tr>
<td>CMTT</td>
<td>4.57 (2.99)</td>
<td>6.00 (1.87)</td>
<td>5.25 (3.40)</td>
<td>7.00</td>
</tr>
<tr>
<td>Beery VMI</td>
<td>12.29 (2.14)</td>
<td>13.20 (.84)</td>
<td>14.50 (2.38)</td>
<td>17.00</td>
</tr>
<tr>
<td>Time 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>df = 4</td>
<td>N = 4</td>
<td>N = 4</td>
<td>N = 4</td>
<td>N = 5</td>
</tr>
<tr>
<td>Mean PAE</td>
<td>12.06 (2.19)</td>
<td>21.17 (5.16)</td>
<td>12.39 (3.50)</td>
<td>13.16 (5.03)</td>
</tr>
<tr>
<td>NKT-A</td>
<td>10.75 (1.89)</td>
<td>10.50 (5.45)</td>
<td>11.00 (1.63)</td>
<td>11.00 (1.87)</td>
</tr>
<tr>
<td>CMTT</td>
<td>8.25 (2.50)</td>
<td>7.75 (1.26)</td>
<td>9.00 (2.94)</td>
<td>8.80 (1.64)</td>
</tr>
<tr>
<td>Beery VMI</td>
<td>14.25 (0.96)</td>
<td>15.00 (3.92)</td>
<td>15.75 (2.22)</td>
<td>14.00 (1.41)</td>
</tr>
</tbody>
</table>

<sup>Note.</sup> <sup>a</sup>Strategy was inferred from patterns of PAE for individual target numbers across trials. Five strategy groups were considered: hook-shaped (hook); tent-shaped (tent); M-shaped (M), low-error (low); and inconsistent error (incon). <sup>b</sup>Not included in calculation of ANOVAs. *p < .05; **p < .01.
Figure 1. Five potential models of best fit applied to the number line estimation task. Models (a) and (b) are predicted from the linear-to-logarithmic shift account. Models (c), (d), and (e) are predicted from the proportion-judgment account. LA = lower anchor (typically 0); Mid = midpoint; UA = upper anchor.
Figure 2. A sample of the variety of play materials available in the classroom, including natural, synthetic, purchased and gathered items: a) coloured wooden rods and wooden trays; b) cork cylinders and plastic metallic pipe connect; c) coloured wooden cubes; d) small wooden trees and wood stain samples; e) patterned tiles and a laminated number grid; f) granite counter top samples; g) carpet samples; h) coloured wooden robots and elephants that connect in a variety of ways; i) a variety of traditional materials including: connecting cubes, pentominos, coloured rubber dinosaurs, and dominoes; j) coloured plastic connecting shapes and a light table; a variety of rocks and shells; coloured wooden pegs and peg board.
Figure 3. Examples of children’s use of play materials: a) trying to completely fill a triangular tray with wooden geometric forms; b) creating a solar system out of pentominos; c) constructing a scene with land beside the ocean (note animals in trees and a stream and stream running through the land mass); constructing a castle scene on a foam board from castle blocks, wooden bungs, and wooden lady bugs; e) building a tower from foam blocks; f) filling in pictures with foam geometric shapes; g) creating a design with a variety of triangles on a peg-board with coloured beads; h) copying a design using coloured wooden blocks inside a wooden tray; i) creating a boat attached to a loading dock with a ramp out of connecting blocks.
Figure 4. Examples of creations built cooperatively during sustained free play.  (a) Children try to build towers of particular heights (e.g., as tall as one child, equivalent heights) using Wedgit™ building blocks.  (b) Children creating a sidewalk demonstrating symmetry.
Figure 5. Graphing activity. Children created graphs of the sum of two rolled dice to determine which sum occurred most often (i.e., which number “won”).
Figure 6. Materials used in the Number Line Estimation task, including (a) the orienting item; (b) the training item; and (c) a sample of a trial.
Figure 7. The lines of best fit for Number Line Estimation patterns of two hypothetical children with the same $R^2_{lin}$. The dotted line represents the equation $y = x$ and the solid line represents the equation $y = 0.25x + 1$. Children with identical $R^2_{lin}$ but with different slopes and y-intercepts can show very different performance.
Figure 8. Number line estimation patterns for individual children (N=21; C1 to C21) at both time points (T1 and T2). Outliers were not included when determining model of best fit. C1-C10 featured here. C11-C21 on the following page.
Figure 8. Continued from previous page.
Figure 9. Group level number line estimation data at Time 1 (T1) and Time 2 (T2). (a) Number line estimation patterns at the group level. (b) Median percent absolute error at the group level.
Figure 10. Mean PAE across target number for groups of individuals with Number Line Estimation patterns categorized in each of the five strategy categories at each time point.
Figure 11. Scatter plots of the relationships between numerical skills (NKT-A) and VMI (Beery VMI) within and between times points.
Figure 12. Varying qualities of hash marks made on the number line estimation task. (a) Child with raw score on Beery VMI of 13. Difference between where the mark starts and where it crosses the number line is 15 mm (at full scale). (b) Child with a raw score of 17 on the Beery VMI. Mark is nearly perfectly straight from where it begins to where it crosses the number line.