Number Line Development of Chilean Children from Preschool to the End of Kindergarten

Chang Xu[1], Sabrina Di Lonardo Burr[2], Heather Douglas[2], María Inés Susperreguy[3], and Jo-Anne LeFevre[1,2]

[1] Department of Psychology, Carleton University

[2] Department of Cognitive Science, Carleton University

[3] Faculty of Education, Pontificia Universidad Católica de Chile

Date of Resubmission: March 2, 2020

Support for this research was provided by the Chilean National Fund of Scientific and Technology Development (CONICYT FONDECYT) through Grants No. 11140899 and 1180675 to María Inés Susperreguy. The authors would like to express their sincere gratitude to the schools, teachers, and children who participated in the study. They are also grateful to the research assistants who contributed to data collection, and to Dr. Andrea Howard for the suggestions on the statistical analyses. Support for data analysis and writing was also provided by the Social Sciences and Humanities Research Council (SSHRC) of Canada through an Insight Grant to J. LeFevre, E. Maloney, H. Osana, and S. Skwarchuk and by Ph.D. scholarships to H. Douglas and S. D. Burr.

Correspondence concerning this research should be directed to Chang Xu at 1125 Colonel by Drive, Ottawa, Canada, or by email at chang_xu@carleton.ca.

This manuscript was accepted for publication in the *Journal of Experimental Child Psychology*. This preprint is the peer-reviewed accepted version but has not yet been copyedited and may differ from the final version published in the journal.

Abstract

Children's performance on number line tasks reflects their developing number system knowledge. Before age five, most children perform poorly on even the simplest number lines (i.e., 0-10). Our goal was to understand how number line skills develop before formal schooling. Chilean preschoolers attempted a 0-10 number line task three times over two years: At the beginning of pre-kindergarten (*M* = 4:7 years:months), at the end of pre-kindergarten (*M* = 5:0), and at the end of kindergarten (*M* = 5:10). We used latent class analysis to group children according to their patterns of performance across number targets. At Time 1, 86% of children had error patterns indicating that they randomly placed estimates on the line. At Time 2, 56% of children continued to respond randomly. At Time 3, 56% showed competent performance across the number line, 23% were accurate only near the endpoints, and 21% were only accurate for low target numbers near the origin. Latent transition analyses showed that the transition from less to more proficient estimation classes was predicted by children's number identification skills. Thus, number line performance changed dramatically from ages 4 to 6 as children began to develop the cognitive and numerical skills necessary to accurately estimate numbers on a number line.

Abstract Word Count: 205

Keywords: preschool children, number line, latent class analysis, Chile, early numeracy, mathematical cognition

Number line estimation is a form of proportional reasoning that is common in educational settings (e.g., Gravemeijer, 2014). Performance on number line estimation tasks improves as estimators gain understanding of the sequential and proportional associations among numbers (e.g., Ashcraft & Moore, 2012; Barth & Paladino, 2011; Laski & Siegler, 2007; LeFevre et al., 2013; Muldoon et al., 2011, 2013). Young children perform poorly on number line tasks before the age of 5 (Cankaya et al., 2014; Berteletti et al., 2010; Whyte & Bull, 2008). Accordingly, this task had poorer validity and classification accuracy in kindergarten compared to a well-established screener measure of early mathematics skills (Sutherland et al., 2021). Furthermore, performance on the task is not correlated with other mathematical skills for children under 6 years of age (see meta-analysis by Schneider et al., 2018). There is limited research on how children approach this task prior to kindergarten (Muldoon et al., 2011; Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009; Xu et al., 2013; Yuan et al., 2019), and no previous studies have investigated the developmental progression in young children's improvement on the number line task in relation to their other cognitive skills. In the present study, we explored the development of Chilean children's number line estimation from preschool to the end of kindergarten using a longitudinal design.

The Development of Number Line Strategies

Performance on the number line task improves with age (Booth & Siegler, 2006; Petitto, 1990; Schneider et al., 2018; Siegler & Booth, 2004). One explanation is that this improvement occurs because children's mental representation of magnitude shifts from a logarithmic to a linear form (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003). This mental magnitude hypothesis is based on observed patterns of number placements, specifically, plotting younger children's estimates as a function of their actual position results in a logarithmic pattern, whereas for older children on the same number line, the pattern of estimates is linear (Ashcraft & Moore, 2012; LeFevre et al., 2013; Siegler & Booth, 2004; Siegler & Opfer, 2003). This shift from logarithmic to linear is also dependent on the range of the number line; the same children may exhibit a linear pattern for a 0-10 number line, but a logarithmic pattern for 0-100.

A second explanation for improvements in number line performance with age is that children learn better strategies, specifically, they use numerical reference points effectively to guide their estimates (Barth & Paladino, 2011; Huber et al., 2014; Link et al., 2014; Peeters et al., 2016; Slusser & Barth, 2017). Children younger than 7 years of age use counting strategies on the number line task. Initially they use a single reference point (i.e., the left endpoint/origin), and count upward, whereas somewhat older children will count from two reference points, the left and right endpoints (Petitto, 1990; Newman & Berger, 1984; Xu & LeFevre, 2016). Accuracy decreases with the number of counts required from either endpoint, resulting in more accurate estimates around those reference points and less accurate estimates as the distance from the reference point increases (Xu & LeFevre, 2016).

Older children and adults also use reference point strategies (Ashcraft & Moore, 2012; Luwel et al., 2018; Peeters et al., 2017; Thompson & Opfer, 2010) but they use more reference points and consider the line more holistically. Reference points can be either explicit (i.e., labelled reference points such as 0 and 100) or implicit (i.e., unlabelled reference points, such as the midpoint). Researchers have inferred strategy use, for example, use of the left endpoint, right endpoint, and midpoint, based on a pattern of error across the number line that is shaped like an "M", with the most accurate estimates (i.e., the least error) around these three reference points (Ashcraft & Moore, 2012; Barth & Paladino, 2011; Newman & Berger, 1984; Petitto, 1990; Xu

& LeFevre, 2016). In contrast, use of the two endpoints but not the midpoint gives a tent-shaped pattern (Ashcraft & Moore, 2012).

Relations between Number Line Performance and Numeracy Skills

The number line task is assumed to reflect children's understanding of the mapping between number symbols and their corresponding quantities (Siegler & Booth, 2004). Children need a variety of skills to accurately estimate on a number line. First, children must have sufficient knowledge of numerical symbols: If they cannot identify the Arabic symbol for "8" then they will not be able to place that number on the number line (Whyte & Bull, 2008). Second, the understanding of the relative magnitude of number symbols is important (Siegler & Ramani, 2008, 2009). For example, children will likely struggle to place numbers on a number line if they do not understand that 9 is less than 10. Third, children also need to know the sequential relations among numbers (e.g., that 8 comes after 7 and before 9) for the range of the number line (Xu & LeFevre, 2016). For example, children who can identify that 8 is closer to 10 than it is to 1, may start at 10 and count backwards to 8 rather than starting at 1 and counting upwards, resulting in more accurate performance.

Magnitude understanding in relation to number symbols is often assessed with the symbolic number comparison task (e.g., which is larger, 4 or 7? Hawes et al., 2019). Kindergarten children (i.e., 5- and 6-year-olds) who could not accurately compare the magnitude of symbolic quantities also had difficulty placing numbers on a 0 – 100 number line (Laski & Siegler, 2007). Similarly, performance on symbolic number comparison was correlated with 0 – 100 number line performance for grade 1 children (i.e., 7-year-olds; Daker & Lyons, 2018). Extending beyond magnitude understanding, ordinal knowledge of the number symbols is also important for number line estimation. Xu and LeFevre (2016) found that preschoolers' performance improved on a 0-10 number line when they received training on the sequential relations between numbers (e.g., what number comes after 4?). Hence, knowledge of the relative magnitudes of number symbols and understanding of the ordinal relations among the number symbols may be important precursors to successful number line estimation.

Relations between Number Line Performance and Domain-General Cognitive Skills

Domain-general cognitive skills such as spatial abilities and executive functions are important skills for number line performance. If children do not understand the spatial relations between numbers (i.e., that the distance between 2 and 3 is the same as the distance between 5 and 6), they cannot accurately place numbers on a number line. For example, 5- to 8-year-old children's ability to determine the spatial distance between placements was related to number line performance (Laski & Siegler, 2007). Furthermore, researchers have found that the spatial arrangement of number board games helped children place estimates on a number line (Siegler & Ramani, 2009; Whyte & Bull, 2008). More specifically, when the board is arranged so that the relative position of the numbers is emphasized, the spatial cues about the relations among numbers may transfer to number line estimation. Moreover, more general spatial abilities are related to estimation performance for children in grades 1 to 3 (e.g., Daker & Lyons, 2018; Gunderson et al., 2012; LeFevre et al., 2013; Xu, 2019). For example, LeFevre et al. (2013) found that spatial abilities (i.e., reasoning and spatial span) predicted growth in number line performance for children in grades 2 to 4. Similarly, Daker and Lyons (2018) found that spatial reasoning predicted number line performance for children in grade 1. These findings indicate that performance on the number line task is related to spatial skills.

Executive functions, such as working memory, inhibition, shifting, and updating, are strongly linked to mathematical performance and learning (see reviews in Blair et al., 2008;

Cortés Pascual et al., 2019). More specifically, executive functions are correlated with number line performance for young children (Friso-van den Bos et al., 2015; Geary et al., 2008; Kolkman et al., 2013; 2014; Laski & Dulaney, 2015; LeFevre et al., 2010). There is evidence that each executive function plays a role in number line performance. First, children rely on working memory to hold and manipulate information relevant to their selected strategy (Kolkman et al., 2014). They need to mentally represent the magnitude of the target number and determine its relation to the number line range and reference points, such as the endpoints or the imagined midpoint. In accord with this view, working memory predicted number line performance for 5- to 7-year-old children (Gimbert et al., 2019). Second, updating is involved on each number line trial, when solvers choose which reference point to use based on the magnitude of the target number. Consistent with this assumption, individual differences in updating predicted which children improved after training on the number line task for 5- and 6-year-old children (Kolkman et al., 2013). Third, selecting a more advanced strategy may require inhibition of a less-advanced, but well-practiced and familiar strategy (Ren et al., 2019). Accordingly, inhibition was related to the rate of improvement in estimation for 5- and 6-year-old children who played numerical board games designed to improve number line estimation (Laski & Dulaney, 2015). In summary, a variety of spatial and executive function skills are correlated with children's performance and their strategy selection on the number line task.

Demographic Factors Related to Number Line Estimation

Demographic factors, such as country of education and family socioeconomic status (SES), are also related to the development of children's number line estimation skills (Ramani & Siegler, 2008; Xu et al., 2013). Compulsory schooling starts as early as age three in some countries and as late as age seven in other countries. Furthermore, children's preschool experiences differ across countries, with some curricula that are largely play based and others that are more academically oriented. For example, children educated in North America show poor performance on the number line task prior to grade 1 (Ramani & Siegler, 2008; Xu & LeFevre, 2016) and also make less accurate estimates than children in China (Laski & Yu, 2014; Siegler & Mu, 2008). Even within a given culture, parental education, educational background, children's intelligence, and family SES are all correlated with numeracy performance (e.g., Baharudin & Luster, 1998; Benavides-Varela et al., 2016; LeFevre et al., 2018; Zippert & Ramani, 2017). In summary, educational, cultural, and socio-economic factors are important for understanding the development of children's number line skills.

Modelling Individual Differences in the Development of Estimation Skill

One of the difficulties in trying to understand how children place estimates on a number line is that accuracy can vary within trials for the same child as well as across children (Bouwmeester & Verkoeijen, 2012; Xu, 2019). Thus, the typical approach of averaging across trials or within grades may obscure relevant variability that can be used to understand children's performance. An alternative approach is to use latent variable analysis, which is based on the assumption that a dataset consists of a mixture of observations, in this case patterns of estimation error, from a number of mutually exclusive latent classes (for categorical variables) or profiles (for continuous variables; Lanza & Cooper, 2016). In essence, this analysis identifies groups of children that show similar performance on the latent variables (Oberski, 2016).

In two studies, researchers used latent variable approaches to differentiate children's patterns of accuracy on a 0-100 number line (Bouwmeester & Verkoeijen, 2012; Xu, 2019). First, Bouwmeester and Verkoeijen (2012) analyzed estimation performance for Dutch children from kindergarten to grade 2 (i.e., aged 5 to 8 years) using latent class analysis. They found that children were clustered into five different classes of number line performance. However, examination of the five classes suggests that there were two main estimation patterns. For two of the classes, the data points were tightly clustered around the estimated linear function, indicating that children were making accurate estimates. For two other classes, however, Bouwmeester and Verkoeijen found that children's estimation patterns were logarithmic-like curves (i.e., children overestimated the location of low target numbers and underestimated the location of high target numbers). Importantly, plots of individual estimates (i.e., Figure 1 in Bouwmeester & Verkoeijen) showed that children in the two logarithmic-like classes had variable patterns of estimation and that the data points were not tightly clustered around the estimated logarithmic function. Finally, three children in kindergarten were classified into a fifth class in which they placed their estimates for all target numbers at approximately "50". Notably, children were told that the midpoint was 50 and thus this pattern reflects that children in this class had a poor understanding of the number line task. In summary, using a latent class approach, Bouwmeester and Verkoeijen identified varying patterns of estimation that occur for young children who are developing their number line estimation skills. However, because they did not present accuracy data for individual target numbers or for individual children, it is difficult to determine how accurately the youngest children were able to estimate on a 0-100 number line in that study.

Second, Xu (2019) used a latent profile approach to analyze the performance of Canadian children in grades 1 and 2 (i.e., aged 6 to 7 years) on a 0-100 number line. For these children, the data showed two profiles that were labelled *uniform* and *variable*. These profiles refer to the uniformity and variability in accuracy across the number line for the two groups. Strategy report data collected by Xu indicated that the children in the variable group used counting strategies that started at either the origin or the right reference point (i.e., 0 or 100); these children often pointed to the number line and recited number sequences. In contrast, children in the uniform group were more likely to use relational strategies (i.e., estimate the location of the target number in relation to various reference points on the line), reflecting their understanding of the ordinal relations among numbers. The findings of Xu and Bouwmeester and Verkoeijen (2012) emphasize the importance of capturing estimation differences across trials and across children. When estimates are collapsed across all trials and all children, individual differences in estimation patterns may be obscured.

The studies described above used latent variable analysis to categorize children into groups at one or more time points. Across the two studies there was evidence that some young children could accurately estimate on the number line (i.e., the linear and uniform patterns) and some could not (i.e., the logarithmic and variable patterns). When longitudinal data are also available, such as in the present study, latent transition analysis can be used to explore the change in patterns of performance over time, and to test whether other numeracy or cognitive skills are related to the probability of transitioning from one group to another. For example, in Xu (2019), grade in school and children's knowledge of number order predicted whether they moved from the variable to the uniform profile over a four-month period, the latter finding presumably reflecting the importance of ordinal knowledge in the development of number line performance. In summary, latent variable analysis can be a useful tool to differentiate estimation patterns for individual children.

The Present Study

In the present study, Chilean children performed a 0-10 number line task three times over a two-year period. They were tested at the beginning and end of pre-kindergarten and at the end of kindergarten. Accuracy of estimates were evaluated by calculating percent absolute error

(PAE). The PAE is the absolute value of the difference between each target and its actual location and thus, lower PAE reflects more accurate estimates. We used these data and latent class analysis (see Analysis Plan for details) to address three research questions.

Question 1. How does number line performance develop before children start primary school?

Based on the limited earlier work on number line estimation with preschool and kindergarten children, we anticipated that many children would find the number line task challenging (Bouwmeester & Verkoeijen, 2012; Xu & LeFevre, 2016). When children have the necessary sequential skills, a uniform estimation pattern emerges (Xu, 2019), such that PAE is similar for all target numbers. However, younger children, such as those in the present study, most often use a counting strategy on the number line task (Petitto, 1990; Newman & Berger, 1984; Xu & LeFevre, 2016). Thus, our first hypothesis was that at the beginning and end of pre-kindergarten, two performance classes would emerge: (a) a *proficient* class, where children would have low PAE across the targets, with the lowest PAEs for targets closest to the endpoints, and (b) a *random* class, where children's PAEs would be high across the entire number line, suggesting random responding (Hypothesis 1a and b; Time 1 and Time 2). Furthermore, by the end of kindergarten, we expected that most children would have an understanding of how numbers from 0 to 10 relate to one another. However, not all children have developed the sequential skills necessary to accurately place estimates on a number line (Xu & LeFevre, 2016). Thus, we anticipated two different classes at Time 3 (Hypothesis 1c): (a) a *proficient* class, and (b) a *variable* class. For both classes we anticipated that children's PAEs would be lowest for numbers that were closest to the endpoints (Bouwmeester & Verkoeijen, 2012; Xu, 2019). However, children in the proficient class were expected to have relatively low PAEs across the rest of the number line, whereas those in the variable class were expected to have higher PAEs for targets that were farther from the endpoints.

Question 2. What factors concurrently relate to number line estimation for children in pre-kindergarten and kindergarten?

Using multinomial logistic regression, we can determine whether SES, verbal counting, number identification, number comparison, and executive function relate to LCA class membership (see Figure 1). Our second hypothesis was that children who had stronger symbolic number knowledge (as measured by the number identification and number comparison tasks) and better executive function skills (as measured by spatial span, digit forward and backward spans, and inhibition tasks) would be more likely to be classified in the proficient class rather than the random or variable classes, controlling for their SES and verbal counting at each of the time points (Hypothesis 2; Daker & Lyons, 2018; Kolkman et al., 2013).

Question 3. What factors contribute to the development of number line estimation skills from pre-kindergarten to the end of kindergarten?

We were interested in which factors would be related to changes in group membership over time (Figure 1). This is a novel use of latent transition analysis with preschoolers performing the number line task. Previous training studies have found that training children on a range of symbolic numerical skills, including number identification and magnitude comparison, led to improvements in number line estimation (e.g., Ramani & Siegler, 2008; Whyte & Bull, 2008). Beyond numerical abilities, various aspects of executive function skills predict improvement in number line estimation (Kolkman, 2013; Laski & Dulaney, 2015). Thus, based on prior work, our third hypothesis was that children's symbolic number knowledge and executive function skills would predict the transition from the *random* to *proficient* class from Time 1 to Time 2 and from Time 2 to Time 3 (Hypothesis 3).

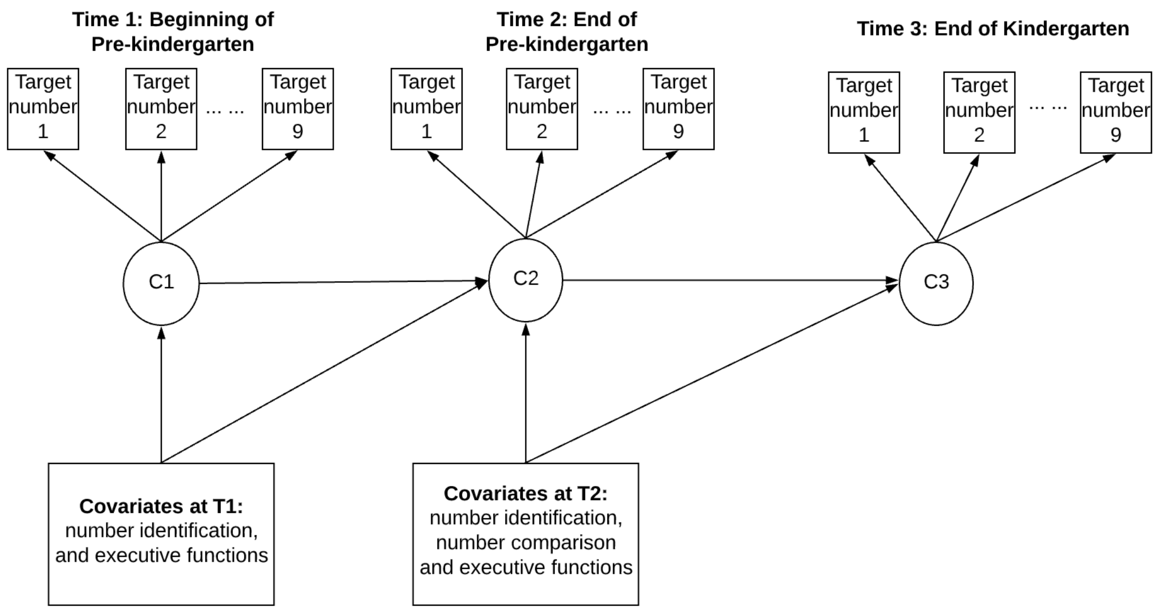

Figure 1. The latent class transition model used for the number line task. C1, C2 and C3 refer to the latent construct extracted from the nine indicators (PAE for target numbers) at Time 1, Time 2, and Time 3, respectively. Covariates were included in the model to determine whether these variables predicted the transitions over time.

Method

The current analyses used data from a longitudinal project examining the development of children's early math skills from pre-kindergarten to the end of kindergarten. In previous papers based on this dataset, we explored the relations between children's home experiences and their early numeracy development (Authors, 2020a, 2020b) and the relations between executive functions and early numeracy performance (Authors, 2019). The present analysis is novel, addressing the specific question of how number line skills develop.

Participants

Children (N = 419; 200 boys) were recruited from pre-kindergarten programs in seven elementary schools in urban Santiago, Chile. In Chile, starting at age four, children attend two years of kindergarten before grade 1 (i.e., pre-kindergarten and kindergarten). Because schooling is strongly linked to family SES in Chile, children were recruited from various urban schools that differentially catered to low-, middle-, or high-SES families. In addition, the parents of the children had a variety of education levels. Children came from a range of SES with 18% of children attending schools serving high/middle-high-SES families, 55% attending schools serving middle-SES families, and 26% attending schools serving low/middle-low-SES families. Levels of parent education reflected this diversity: 26% of parents reported their highest level of education as a high school diploma or less whereas 33% of parents reported completing university degrees. SES status of the school (i.e., low, medium, or high) was highly correlated with parental education, $r(390) = .616$, $p < .001$. Thus, the sample was culturally homogenous but heterogenous with respect to SES and parent education.

Following ethics approval from (blinded), school principals were contacted. Information was shared with parents of pre-kindergarten children at participating schools. Interested parents provided informed consent for their child to participate and children provided oral assent prior to each testing session. The school year in Chile starts in March and ends in December. Children were tested at three time points. Time 1 was the early fall term (March – April of 2016) at the beginning of pre-kindergarten. Time 2 was approximately seven to eight months later, at the end of the second term (October – November of 2016), and Time 3 was approximately nine to ten months later, near the end of the kindergarten second term (September – October of 2017). Testing at each time point involved two 25-minute sessions and was conducted one-on-one with the child in a quiet area of the school. Attrition was low. Of the original 419 children at Time 1, 13 did not participate in Time 2 testing and 51 did not participate in Time 3 testing. At Time 1, children had a mean age of 4:7 (years:months, range 3:4 to 5:7; $N = 419$). At Time 2, children had a mean age of 5:0 (range 3:10 to 5:10; $N = 406$). At Time 3, children had a mean age of 5:10 (range 4:9 to 6:6; $N = 368$).

Materials and Procedure

Only those measures used in the current analyses are described. Children also completed other mathematics and vocabulary tasks. A complete description of measures can be found in Authors (2020a).

Executive Function

Four measures of executive functioning were used. Executive functioning tasks were completed twice, at Time 1 and Time 2.

Spatial span. The *PathSpan* task is a measure of visual-spatial short-term memory (Alloway et al., 2008). Children see nine green dots on the tablet screen (Hume & Hume, 2014). Dots were arranged unsystematically, however, the arrangement was the same on each trial. The dots light up sequentially in varying patterns and children attempt to reproduce the pattern by tapping on the dots in the same order. To ensure children understand the task, they practice with a 2-dot sequence. Testing begins with a 3-dot sequence. Children are given three trials for each sequence (span) length and testing is discontinued when the child is incorrect on all three trials for a given span length. Scoring was the total number of correctly reproduced sequences. Test reliability comparing sub-scores for the first, second and third trials of each span was acceptable at both Time 1 and Time 2, Cronbach's α of .75 and .79, respectively.

Forward and backward digit span. The forward digit span task is a measure of verbal short-term memory and the backward digit span task is a measure of working memory (Alloway et al., 2008). The tester recites a number sequence and the child repeats the sequence in order (forward digit span) or backward (backward digit span). Children begin with a 2-digit span practice trial (e.g., tester say 9-1, child repeats 9-1 for forward digit span or 1-9 for backward digit span). There are four trials for each span length and testing is discontinued when a child incorrectly repeats all trials for a given span. The maximum span length was five digits for both forward and backward span tasks. Scoring was the total number of sequences correctly repeated and thus could range from 0 to 16. Test reliability comparing sub-scores for the first, second, third, and fourth trials of each span was good for both forward digit span, Cronbach's α of .88 (Time 1) and .84 (Time 2), and backward digit span, Cronbach's α of .88 (Time 1) and .89 (Time 2).

Head-Toes-Knees-Shoulders. This HTKS task (McClelland et al., 2014) is a measure of behavioural self-regulation and inhibition. In the first part, there are four paired rules and the instructions match the rule (e.g., touch your head when told to "touch your head"). In the second part, there are four new paired rules (e.g., touch your head when told to "touch your toes"). If the child correctly responds to the four rules a new set of rules is introduced (e.g., touch your head when told to "touch your knees"). In the second part, there are a total of 20 trials. Trials are scored as follows: 0 for an incorrect response, 1 for a self-corrected response, and 2 for a correct response. Maximum possible score is 40. Reliability for both Time 1 and Time 2 was high, with Cronbach's α of .95 and .92, respectively.

Math Measures

Verbal counting. In this task, children were asked to count as high as they could starting from the number one. Scoring was the highest number children counted to before making a mistake. Children completed this task at Time 1 and Time 2. Test-retest reliability across Time 1 and Time 2 scores was low (Cronbach's α = .57) suggesting that children's performance improved differentially over time.

Number identification. In this task (adapted from Purpura & Ganley, 2014) children were shown numbers on flashcards and then responded by naming the digit. The number identification task was more advanced for each time point. At Time 1, there were 17 trials: All nine 1-digit numbers and eight 2-digit numbers. At Time 2 there were 18 trials: Five 1-digit numbers, eight 2-digit numbers, three 3-digit numbers, and two 4-digit numbers. At Time 3 there were 26 trials: All nine 1-digit numbers, twelve 2-digit numbers and the 3- and 4-digit numbers used at Time 2. Testing was discontinued after three consecutive incorrect responses for Times 2 and 3; there was no discontinue rule for Time 1. Scoring was the total number of correct responses. Task reliability was based on all 17 trials at Time 1, and on the comparable set of trials (i.e., trials including two-digit numbers in the twenties) at Time 2 (9 items) and at Time 3 (17 items). Notably, because of the discontinue rule, 48% of children at Time 2, and 59% of children at Time 3 completed the comparable set of items. Reliability was better at Time 1 than at Time 2 and Time 3 (Cronbach's α = .89, .62, and .66, respectively), reflecting differences in the number of comparable trials, differences in the discontinue rule, and variability of the range of numbers at each time point.

Number comparison. The symbolic version of the numeracy screener task was used to access number comparison (see task details in Nosworthy et al., 2013 and Hawes et al., 2019). Children see pairs of single-digit numbers and are required to put a line through the larger digit of each pair. There were three example items and nine practice items. Children had a maximum of two minutes to cross out the larger number of 56 pairs. Scoring was the number of correct responses divided by the time taken to complete the task in seconds. Children completed this task at Time 2 and Time 3. Test-retest reliability across the two years was good (Cronbach's α = .76).

Number line estimation. In the *EstimationLine* task, children see a number line with the endpoints 0 and 10 on the tablet screen (Hume & Hume, 2014). A target number is shown on the screen and children tap a spot on the number line where they think the target number is located. To familiarize children with the task, they begin with three practice trials where they tap a marked spot on the number line. Following the practice trials, children complete nine randomly ordered trials, one for each digit (1 to 9). Percent absolute error (PAE) between the placement of the number and the actual location of the number was calculated for each trial. Reliability comparing PAE across the nine trials improved over time (Time 1, Cronbach's α = .69; Time 2, Cronbach's α = .72; Time 3, Cronbach's α = .80).

Missing Data

The cross-sectional latent class analysis (LCA) at each time point was conducted on data from all of the children who completed the number line task (388, 403, and 368, for Times 1, 2, and 3, respectively). Longitudinal analyses were conducted using the data from all children, based on the full information maximum likelihood estimation (FIML) in conjunction with robust maximum likelihood estimation to handle missing data (Enders, 2010), rather than a list-wise deletion strategy focusing only on children having completed the number line tasks across all three time points. FIML estimation does not replace the missing values, rather, it estimates model parameters based on all of the variable information in the variance-covariance matrix. This approach has been shown to result in unbiased parameter estimates under even a high level of missing data for longitudinal analyses under missing at random assumptions (Enders, 2010). Furthermore, to ensure that the results from the longitudinal latent transition analysis (LTA) model were unbiased by this missing data approach, the results from the cross-sectional LCA conducted on the time-specific samples were examined to determine whether the results converged with those from the longitudinal LTA.

Results

In the present study, children's accuracy on the number line task was indexed with PAE, which captures the absolute value of the difference between the target and the actual location by taking into account the scale of the number line. Thus, lower PAE reflects more accurate estimates. The PAEs for the nine target numbers of the 0-10 number line task were used in a latent class analysis as the indicators of children's estimation classes. Models were fit in several steps to evaluate the number of classes (i.e., groups of children) that best captured the data at each time point.

Descriptive Statistics

Table 1 presents the means, standard deviations, range, and skewness for the raw scores of the number line task and the predictors used in the present study. As shown in Table 1, the mean PAE for each target number on the number line task ranged from 26% to 41%, suggesting that, at the beginning of pre-kindergarten, many children had difficulty with the task. At Time 2, mean PAEs for each target number ranged from 19% to 29%, suggesting that at the end of pre-kindergarten, many children still had difficulty doing the number line task. In contrast, by Time 3, most of the mean PAEs for each target number were less than 20% (ranging from 10% to 22%), suggesting that at the end of kindergarten, many children could do the task. Overall, mean PAEs were 31%, 24%, and 17%, at Times 1, 2, and 3, respectively. Similarly, in other studies of preschool and kindergarten children’s performance on typically bounded (i.e., 0-10 or 0-100) number lines, mean PAEs ranged from 17% - 26% (Muldoon et al., 2011; Praet & Desoete, 2014; Ramani & Siegler, 2008; Siegler & Ramani, 2008, 2009; Xu et al., 2013). These results are consistent with the view that 4- and 5-year-old children generally show poor performance on the number line task.

Table 2 presents the correlations among the raw scores of the variables at Time 1, Time 2 and Time 3. Results showed that all of the variables were significantly correlated with each other at both time points with one exception: Number line estimation at Time 1 and at Time 3 were not significantly correlated, reflecting the unsystematic nature of responses at Time 1. In general, correlations were higher within each timepoint than between timepoints.

Data Reduction

Principal component analyses (PCAs) were conducted for the executive function measures to create component scores reflecting the shared variance among the measures. In particular, at Time 1, the PCA resulted in one component (factor loadings for spatial span, inhibition, digit forward span, and digit backward span = .64, .78, .69, and .80, respectively), accounting for 53% of the variance in these measures. Similarly, at Time 2, the PCA resulted in one component (factor loadings for spatial span, inhibition, digit forward span, and digit backward span = .69, .71, .80, and .71, respectively), accounting for 53% of the variance in these measures. These two executive component scores were saved and used as predictors in subsequent analyses.

Table 1

Descriptive Statistics of the Raw Scores for the Number Line Task (Percentage Absolute Error) and Covariates

Variables	Time 1 (*N*=388)				Time 2 (*N*=403)				Time 3 (*N*=368)			
	M	*SD*	*Range*	Z_{skew}	*M*	*SD*	*Range*	Z_{skew}	*M*	*SD*	*Range*	Z_{skew}
Number Line Task												
Number 1	34.55	32.68	89.70	5.71	20.95	27.05	89.90	13.77	10.00	15.97	90.00	32.51
Number 2	28.46	27.13	79.80	7.13	19.34	21.64	79.80	15.40	12.42	13.77	80.00	30.24
Number 3	26.38	21.15	69.70	6.27	19.03	16.99	69.80	13.30	16.42	13.17	70.00	19.64
Number 4	26.20	18.41	60.00	2.86	21.73	15.72	59.80	6.48	18.84	12.09	59.90	9.30
Number 5	26.46	15.90	59.90	.50	24.25	15.09	50.00	2.03	21.31	12.07	50.00	3.68
Number 6	29.43	17.08	69.80	1.33	25.20	15.93	59.40	4.10	22.47	13.91	60.00	2.87
Number 7	31.28	19.83	69.80	2.88	25.40	19.50	69.80	6.46	20.08	16.56	69.80	7.91
Number 8	35.77	24.93	80.00	2.60	26.70	21.91	79.80	7.69	18.19	18.26	80.00	11.65
Number 9	41.09	30.61	90.00	1.31	29.33	28.68	90.00	6.19	15.83	20.81	90.00	14.06
Mean PAE numbers	30.94	12.67	69.70	1.28	23.57	12.80	60.00	5.19	17.28	9.63	53.30	10.47
Covariates												
Spatial span[1]	3.19	2.10	11	8.76	4.79	2.67	12	5.13	-	-	-	-
Inhibition (HTSK)[1]	18.61	12.93	40	-1.06	24.50	10.68	40	-5.98	-	-	-	-
Digit forward span[1]	6.42	2.79	15	-0.64	7.88	2.52	16	2.28	-	-	-	-
Digit backward span[1]	0.68	1.43	7	16.80	2.16	2.20	10	4.45	-	-	-	-
Verbal counting[2]	13.18	8.47	69	15.13	26.88	18.74	99	18.58	-	-	-	-
Number identification[1]	6.21	4.16	17	4.55	5.65	2.84	16	6.96	14.85	5.19	26	-0.59
Number comparison[3]	-	-	-	-	0.22	0.10	0.48	-0.38	0.32	0.10	0.61	-1.46

Note. Z_{skew} > 3.29 corresponds with an alpha level 0.05, representing that the distribution of the sample is non-normal; [1] total correct; [2]highest count; [3]items correct per second.

Table 2

Correlations Among variables at Time 1, Time 2 and Time 3

	Time 1				Time 2					Time 3	
	1	2	3	4	5	6	7	8	9	10	11
1. Executive Function T1	-										
2. Verbal Counting T1	**.55**	-									
3. Number Identification T1	**.56**	**.62**	-								
4. Number Line T1	**-.26**	**-.30**	**-.34**	-							
5. Executive Function T2	**.64**	**.42**	**.48**	**-.22**	-						
6. Verbal Counting T2	**.45**	**.53**	**.57**	**-.21**	**.43**	-					
7. Number Identification T2	**.44**	**.48**	**.67**	**-.26**	**.47**	**.74**	-				
8. Number Comparison T2	**.36**	**.34**	**.36**	*-.15*	**.33**	**.36**	**.40**	-			
9. Number Line T2	**-.20**	**-.19**	**-.30**	*.13*	**-.29**	**-.19**	**-.26**	**-.30**	-		
10. Number Identification T3	**.47**	**.49**	**.65**	**-.22**	**.49**	**.60**	**.77**	**.39**	**-.31**	-	
11. Number Comparison T3	**.38**	**.30**	**.37**	*-.13*	**.32**	**.39**	**.42**	**.61**	**-.25**	**.49**	-
12. Number Line T3	*-.16*	*-.15*	**-.24**	.07	**-.19**	*-.16*	**-.23**	*-.16*	**.23**	**-.33**	**-.30**

Note. Bolded numbers were significant at $p < .001$; italicized numbers were significant at $p < .05$.

Latent Class Analyses

Analysis Plan

For the latent class analyses, PAEs on each of the target number were recoded into categorical variables using a conservative cut-off value of 20% PAE. All PAEs less than or equal to 20% were coded as 1 (i.e., likely to be legitimate estimates) and all PAEs greater than 20% were coded as 0 (i.e., likely to be random estimates). Thus, PAE was recoded as a binary variable (i.e., 0 or 1). This cut-off value of 20% was chosen for two reasons. First, the existing literature on preschoolers and kindergarteners has reported mean PAEs ranging from 17% to 26% (Berteletti et al., 2010; Praet & Desoete, 2014; Ramani et al., 2017; Sasanguie et al., 2012; Xu & LeFevre, 2016; Xu et al., 2013). Thus, performance below 20% PAE was assumed to reflect a better understanding of the task than performance above 20%. Second, if a child guessed completely at random, then for each of the nine trials they would produce an estimate that falls between 0 and 10. If this was repeated with a large number of children who all guessed at random, eventually the mean estimate for each trial would be placed at "5". Thus, the average PAE for targets 1 through 9 would be 40%, 30%, 20%, 10%, 0%, 10%, 20%, 30%, and 40%, respectively, which would result in a U-shaped curve. The overall expected PAE for children guessing at random would be 22.2%. These recoded binary scores were used to classify children in the subsequent latent class models. Reliability based on the binary scores improved from Time 1 (Cronbach's α = .60) to Time 2 and 3 (Cronbach's α = .72 and .71, respectively). Latent class analysis (LCA) with binary scores was selected over latent profile analysis (LPA) with the full range of values because LPA is extremely sensitive to non-normal data (Bauer & Curran, 2003). Inspection of the raw data showed that a significant portion of children placed estimates unsystematically, resulting in wildly varying estimates. As a result, the profile differences in

LPA appear to largely reflect sensitivity to the variability in the estimation patterns that are uninterpretable, and this violation of model assumptions (e.g., effects of skewness and kurtosis) may lead to an overextraction of profiles and inaccurate model parameter estimates (Bauer & Curran, 2003). Given these issues with LPA, LCA was selected in the present study.

A cross-sectional LCA at each time point was used to determine whether similar estimation classes emerged. Five different indicators of model fit were used because there is no single statistical indicator commonly agreed on for use in determining the appropriate number of classes in mixture models (Nylund et al., 2007). Each initial model included a single class. Subsequent models then were compared against the previous models to determine the number of latent classes that best fit the data at each time point. Fit indices for each model were compared to select the best fitting model among those considered. Specifically, five indices were considered: Scaled log-likelihood value (LL; higher values indicate better fit); Bayesian Information Criterion (BIC; look for last relatively large decrease in BIC; Nylund-Gibson et al., 2014); Lo-Mendell-Rubin Likelihood ratio test and the bootstrap likelihood ratio test (LMR-LRT and BLRT; $p < .05$ suggests n classes is better than n -1 classes; Nylund et al., 2007).

Each model was tested with multiple sets of random start values exceeding 1000, with 50 initial stage iterations (Geiser, 2013). The best log likelihood value was replicated, suggesting that the optimal set of parameter estimates is trustworthy. We started with a one-class model, and then increased the number of classes one at a time. Because the first non-significant p value for the LMR-LRT or BLRT occurred with the five-class model at each time point, indicating that adding more classes to the model would not statistically improve model fit, we stopped testing additional models at the five-class model.

Fit indices for each model are reported in Table 3. Together, the fit statistics did not mutually identify a single LCA model at any of the time points as the best fitting model. It is common that fit indices do not support the same model, and interpretability for the potential models at each time point should be examined to inform model selection (Nylund-Gibson & Choi, 2018). Accordingly, we examined the item plots (see Figure 2) and the spaghetti plots (see Figure 3) for further information about children's pattern of estimation within each class. Although these spaghetti plots are noisy, they show that there are consistencies across individual children. In contrast, the item plots, which are based on averaged PAEs for each target number for each class, more clearly highlight the overall pattern of performance. To select the best model, we considered whether the pattern of results for each step of the analysis resulted in conceptually and statistically superior solutions compared to the previous step. When fit indices were similar for models with different numbers of classes, and the additional class did not improve the model from a theoretical standpoint, parsimony was favoured and the model with fewer classes was chosen. After the final models were selected, we examined the entropy values as an evaluation tool for overall classification of the model: Entropy values > .70 suggests well-differentiated classes (Morgan, 2015; Nagin, 2005). The entropy values for all chosen models were good, suggesting that the latent classes are clearly discriminable.

Question 1. How does number line performance develop before children start primary school?

Cross-sectional Latent Class Analyses. As shown in Figure 2 (a and b), similar two-class formations at Time 1 and Time 2 were observed. For children in the proficient class, estimation errors were relatively stable, although there was a slight increase in error with the size of the target number (i.e., the number of counts from the left endpoint). In contrast, in the random class, estimates had a U-shaped pattern, with the highest PAEs for the endpoints, suggesting that responses were unrelated to target magnitudes. Examination of the three- and four-class models showed that the random class split into multiple classes with differences in performance that were not easily interpretable. Thus, the two-class models were chosen as the final LCA models at Time 1 and Time 2.

At Time 3, we did not observe the U-shaped pattern expected if children were randomly placing estimates. Moreover, contrary to our hypothesis that there would be a two-class model at Time 3 (proficient and variable; Hypothesis 1c), a three-class model best captured the heterogeneity in children's estimation patterns. As shown in Figure 2c, in the proficient class, estimation errors were low and performance was very similar to that in the proficient classes at Times 1 and 2, although children had slightly lower error at both endpoints, not just the origin. In contrast, children in the two variable classes at Time 3 showed higher estimation errors than those in the proficient class. These two variable classes were named the *origin* and *two-endpoints* classes.

For children in the origin class, estimation errors increased with the number of counts and thus performance was only accurate for target numbers nearest the origin (i.e., numbers 1, 2, and 3). Percentage of error for larger target numbers (i.e., 7, 8, 9) was greater than 30%, indicating that children had difficulty estimating numbers far from the origin. For children in the two-endpoints class, performance was only accurate for targets near either endpoint (i.e., numbers 1, 2, 8, and 9). The proficient classes at Times 1 and 2 also showed increasing errors as target size increased. However, their estimates for the largest numbers were lower than 20% and their PAEs for all targets were relatively low in comparison to those of the other two classes. Despite the somewhat uneven performance of the origin and two-endpoints classes, these children were responding accurately to some of the targets. Thus, we conclude that most children had some understanding of the number line task at Time 3.

Table 3

Fit Statistics for LCA Models with 1-5 Classes at Time 1 (N=388), Time 2 (N=403), and Time 3 (N=368).

Class-solutions	LL[a]	BIC[b]	LMR-LRT[c]	BLRT[d]	Entropy
Time 1					
1	1.00	4714	-	-	-
2	.99	**4535**	$p < .001$	$p < .001$	.907
3	**1.10**	4463	$p = .001$	$p < .001$	.832
4	1.09	4462	***p* = .009**	***p* < .001**	.730
5	1.08	4502	$p = .348$	$p = .667$	.738
Time 2					
1	1.00	4737	-	-	-
2	1.05	**4317**	$p < .001$	$p < .001$	.806
3	1.10	4278	***p* = .006**	$p < .001$	.800
4	**1.19**	4278	$p = .271$	***p* < .001**	.794
5	1.02	4311	$p = .031$	$p = .071$	.855
Time 3					
1	1.00	3659	-	-	-
2	1.07	**3361**	***p* < .001**	$p < .001$	.749
3	**1.27**	3308	$p = .285$	$p < .001$	.775
4	1.09	3307	$p = .130$	$p < .001$	.830
5	1.07	3328	$p = .088$	$p < .001$	.845

[a] LL (Scaling correction factor for MLR); [b] BIC (Bayesian information criterion); [c] LMR-LRT (Lo-Mendell-Rubin likelihood ratio test); [d] BLRT (bootstrap likelihood ratio test). The Bolded values indicate the “better” model.

Interpretation of Error Patterns. Although LCA shows patterns of estimates for each target, the patterns still reflect the mean estimates for each target for each class. Spaghetti plots, in which each line represents a child's PAE for each target number, are shown in Figure 3. These plots highlight both variability across individuals and areas of commonality. The denser (i.e., darker) areas reflect common placements of estimates across children whereas the less dense areas reflect less common placements. Both mean patterns and spaghetti plots were used to infer the strategies children may have used in each of the classes, as described below.

Proficient Classes. At Time 1 (Figure 3a), Time 2 (Figure 3b), and Time 3 (Figure 3c), for the proficient group, the densest areas of the spaghetti plots are flat along the bottom of the graph, slightly increasing as target numbers increase at Times 1 and 2. The dense portions are even darker at Times 2 and 3 because there are more children in these classes than at Time 1. For all target numbers, the majority of children in these classes (i.e., > 50%) are making relatively accurate estimates (i.e., PAE is less than 15%). The individual lines show that not all children in this group make accurate estimates across the whole line, because some children have high PAE for large target numbers. However, the low PAE for low target numbers suggests that children in the proficient classes at all three time points understood how to approach the number line estimation task, even though they were not uniformly accurate across the range.

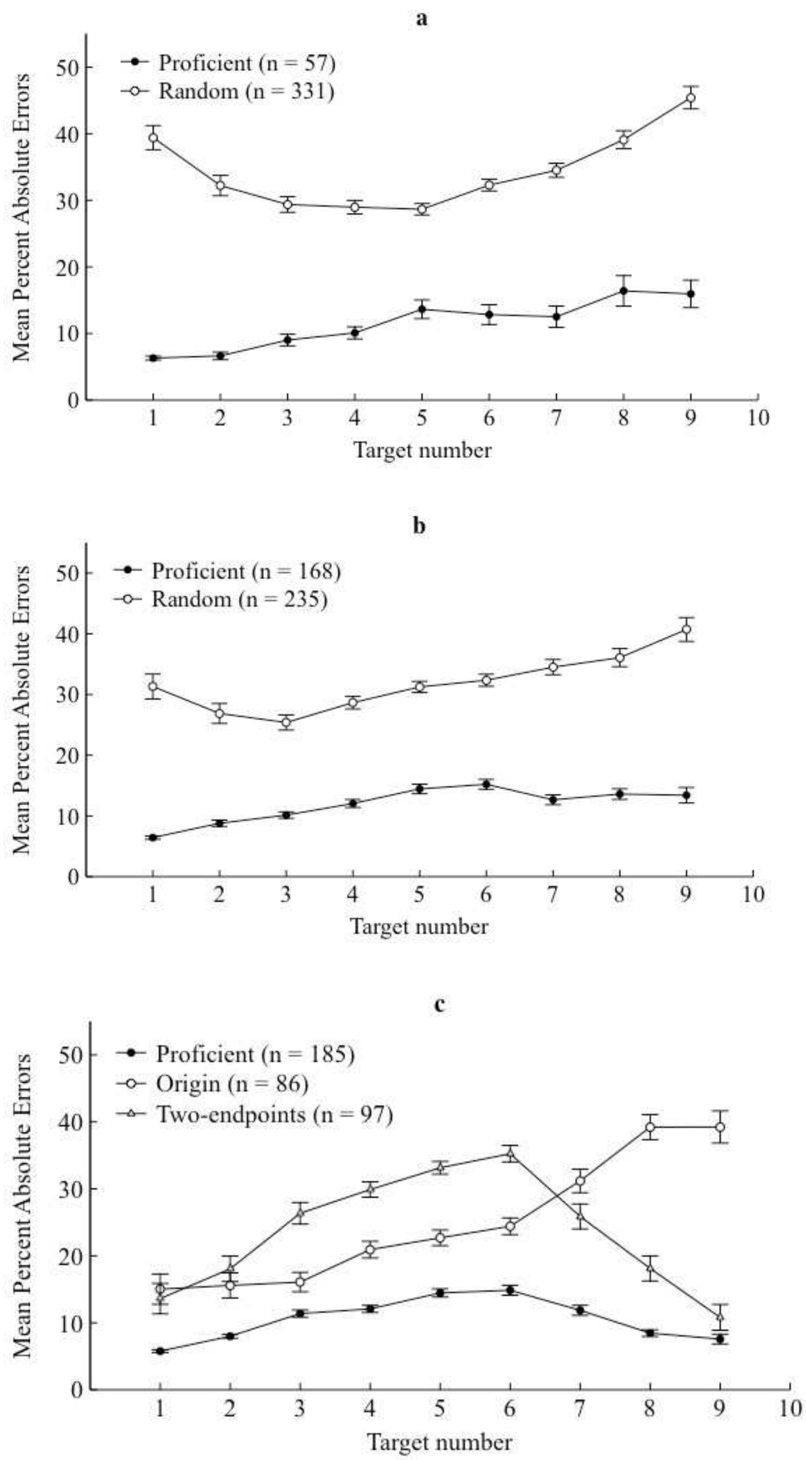

Figure 2. Mean percent absolute errors (PAE) for the two-class LCA models at Time 1 (a) and Time 2 (b), and the three-class LCA model at Time 3 (c). Error bars represent the standard errors of the means for each estimated target.

Most importantly, there is no evidence for a U-shaped pattern in the proficient classes, suggesting that these children were not randomly placing estimates. Instead, at Times 2 and 3, there is some evidence of a dark *inverted* U-shape in the proficient classes, with more accurate estimates around the endpoints than the middle. This pattern suggests that some children in the proficient group may be counting from both endpoints. At Time 3 (Figure 3c), there are fewer extremely high PAEs among the children in the proficient class than at Times 1 and 2. In general, by Time 3, children in the proficient class appear to have developed both the numerical and spatial knowledge required to successfully estimate on the 0-10 number line, as reflected by their relatively low PAEs across all target numbers.

Random Classes. In contrast, the error patterns in the random classes are very different than those in the proficient classes at Times 1 and 2. Overall, the density patterns shown in the spaghetti plots are U-shaped, reflecting poorer performance near both endpoints than in the middle of the line. If estimates on the 0 to 10 number line were random, then for each target number (1-9), the mean estimate would be 5. Accordingly, fully random responding would produce the highest error around the low and high numbers and the lowest error around the middle numbers, consistent with the averaged patterns shown in Figure 2. In the spaghetti plots shown in Figures 3a and 3b, this pattern of responding is more pronounced than in Figure 2, but consistent with the random choice of locations for many children on many trials.

Endpoint Classes. At Time 3 there is no random class. Instead, we observed two classes in addition to the proficient class, the origin and the two-endpoints classes. Similar to the proficient class, children in the two-endpoints class show an inverted U-shaped pattern such that the estimates for targets in the middle of the line are less accurate than those at the endpoints. However, the estimates of children in the two-endpoints class in the middle of the line are less accurate than those of the proficient group. Overall, this pattern suggests that children in the two-endpoints class are probably using both endpoints, counting either upwards or downwards depending on the magnitude of the target. Accuracy decreases farther from the endpoints, suggesting that their increments are not equal in size. Similarly, in the origin class, error increases linearly such that estimates are less accurate as target numbers increase, suggesting that children are counting from the origin. As in the two-endpoints class, children’s increments may not be equal in size and thus, error increases linearly as the targets increase, presumably because children do not use the upper endpoint to calibrate their estimates.

Additional analyses on the statistical fit (i.e., linear, logarithmic, power, and one-cycle comparisons) of each child’s response in relation to the LCA classifications were conducted. These statistical models did not provide further insights into the types of strategies children may be using and thus are presented in the Supplementary Material.

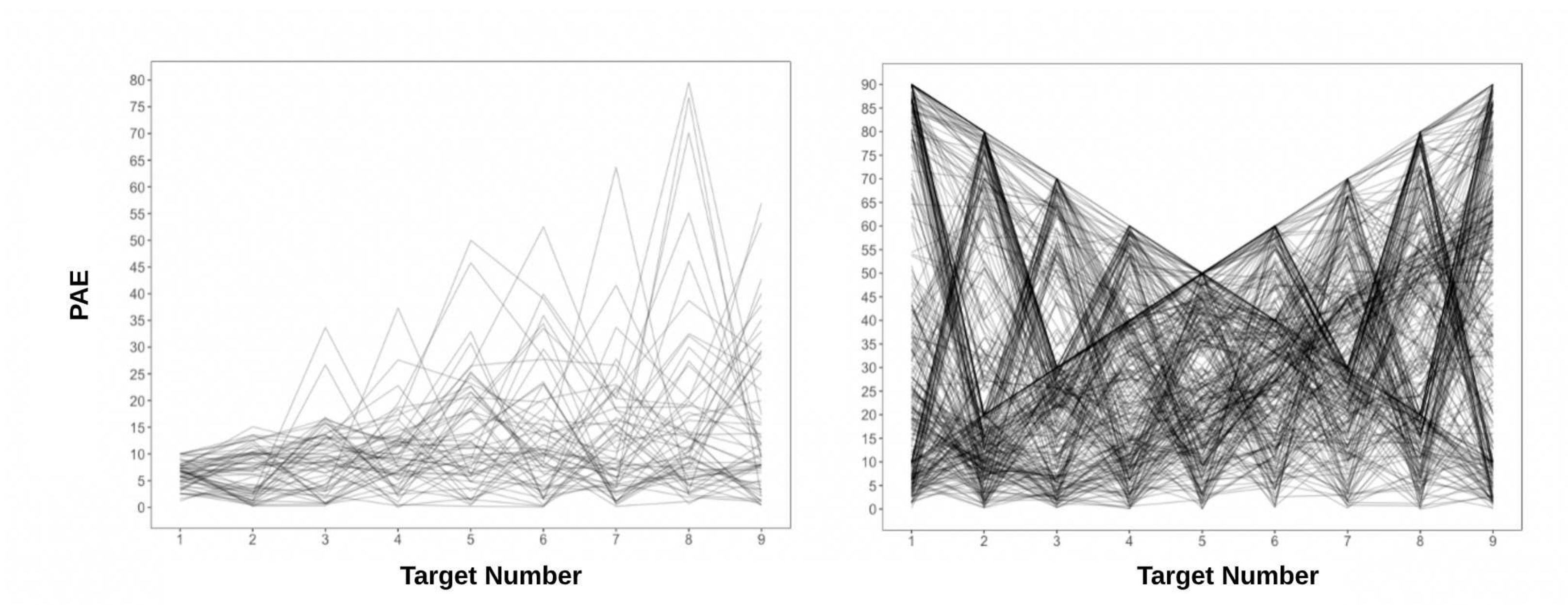

Figure 3a. Spaghetti plots for proficient (left) and random (right) classes at Time 1. Each line connects an individual child's PAE for each target number. Darker areas represent more similar estimates across children and thus show patterns of performance for the group.

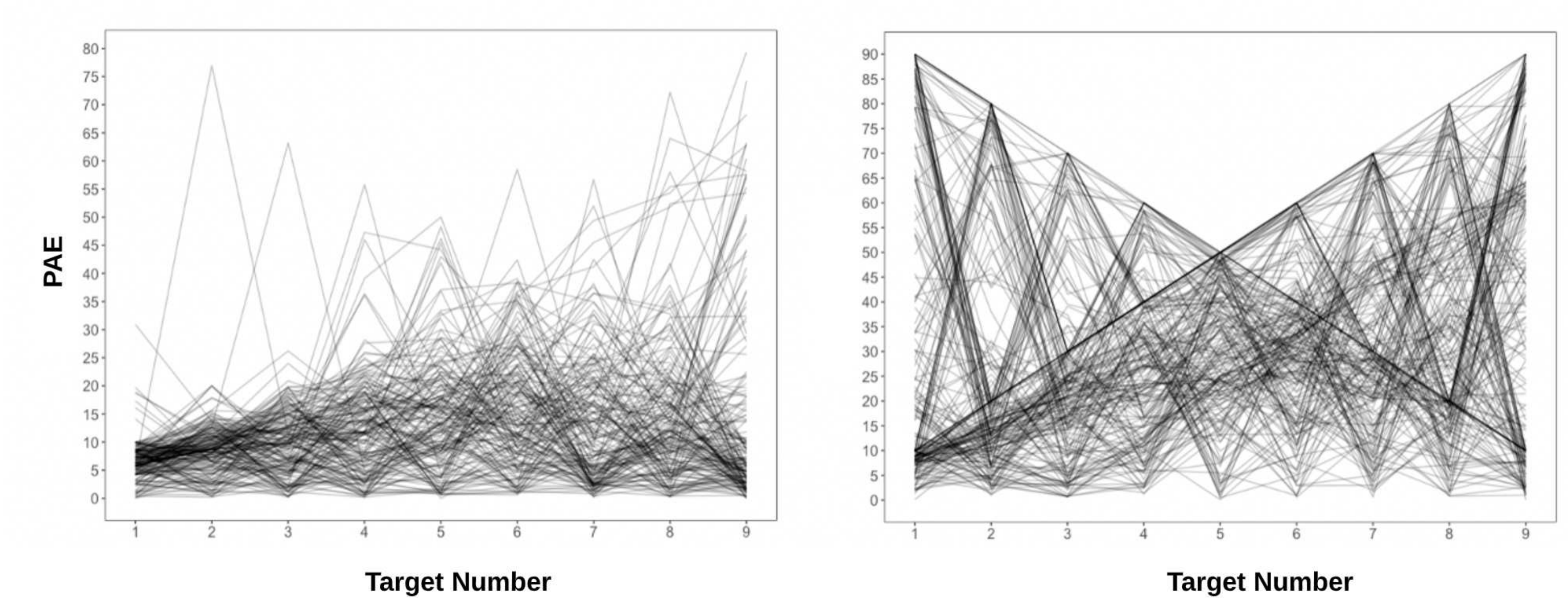

Figure 3b. Spaghetti plots for children in the proficient (left) and random (right) classes at Time 2. Each line connects an individual child's PAE for each target number. Darker areas represent more similar estimates across children and thus show patterns of performance for the group.

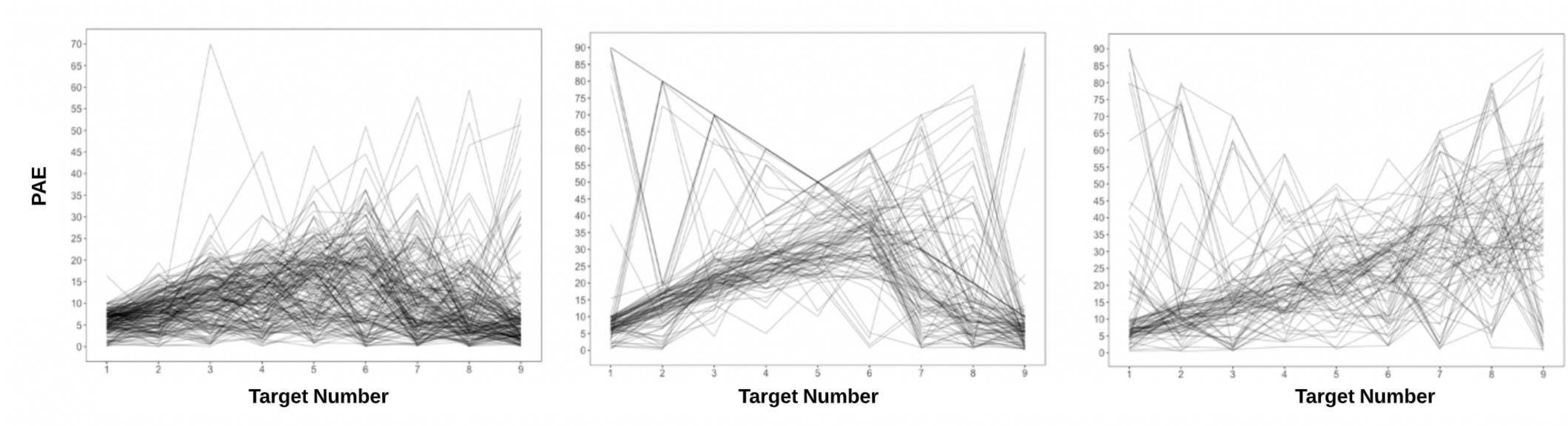

Figure 3c. Spaghetti plots for children in the proficient (left), two-endpoints (center) and origin (right) classes at Time 3. Each line connects an individual child's PAE for each target number. Darker areas represent more similar estimates across children and thus show patterns of performance for the group.

Latent Transition Analysis (LTA). An LTA was conducted based on the chosen LCA models from Step 1. Given that the structure and number of classes (i.e., two classes; proficient and random) extracted from the LCA models at Time 1 and Time 2 was similar, whereas different classes emerged at Time 3 (i.e., three classes; proficient, origin, and two-endpoints), a partial measurement invariance model was tested, with all measurement parameters held equal between Time 1 and Time 2, and freely estimated at Time 3.

The results of the unconditional LTA model are highly similar to the results from the LCA (*LL* = 1.29, *BIC* = 12108, Entropy = .76). At Time 1, the majority of the children were classified in the random class, whereas very few children were classified in the proficient class (see Figure 4). At Time 2, the percentage of children in the proficient class had increased, but more than half of the children were still classified in the random class. By Time 3, three different class formations emerged (proficient, two-endpoints, and origin classes) showing considerable improvement in children's understanding. Most children showed performance that reflected some understanding of number relations but those in the two-endpoints and origin classes applied strategies that were highly dependent on the relation of the target number to the endpoints of the number line.

Figure 4 presents the patterns of change of children from Time 1 to 3 based on the unconditional model. As shown in Figure 4, from Time 1 to Time 2, many children stayed in the random class or stayed in the proficient class; however, about one-third of the children improved in that they transitioned from the random to proficient class. In contrast, few showed "worse" performance, that is, transitioned from the proficient class to the random class. As anticipated, many children found the number line task very difficult at this point in their development and 56% still showed random performance at Time 2. By Time 3, 56% of children were classified in the proficient class. Most of those children were also in the proficient class at Time 2. Of the children in the random class at Time 2, many transitioned to the proficient class at Time 3; fewer transitioned to the two-endpoints or the origin classes. Overall, children were more likely to remain in the proficient class, or transition from the random to the proficient class, than to move to the two-endpoints or the origin classes. Of interest, therefore, was whether other information about the children is helpful in predicting transitions from one class to another across time.

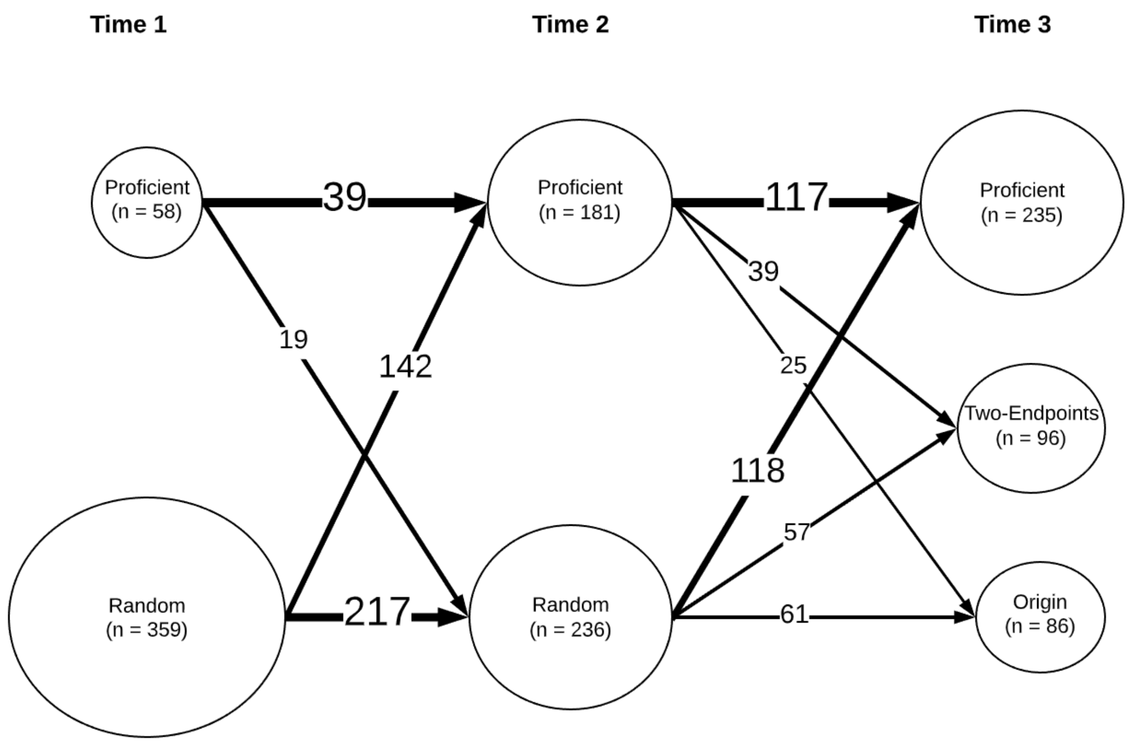

Figure 4. Transition patterns from Time 1 to Time 3 in the unconditional LTA model ($n = 417$). The size of the circles is proportional to the number of children in each group at each time point. The thickness of the arrows approximates the proportion of children and the numbers on the arrows represents the number of children who transitioned from each class at each time point.

Question 2. What factors concurrently relate to number line estimation for children in pre-kindergarten and kindergarten?

In a conditional LTA model, when examining the concurrent relations between the predictors and class membership, the model considers the predictors as additional indicators of class membership, which may lead to overextraction of classes (Nylund-Gibson & Masyn, 2016). Thus, in the present research, we were only interested in how the predictors were related to concurrent number line estimation, not how the predictors acted as indicators of class membership. For parsimony, multinomial logistic regression was conducted to examine whether each of the predictors was related to the classification of children into classes at each of the time points of the unconditional LTA model. This analysis focused on concurrent relations between predictors (i.e., socio-economic status, verbal counting, number identification, number comparison and executive functions) and class membership at Times 1, 2 and 3 (note that number comparison was not measured at Time 1 and executive functions were not measured at Time 3). As shown in Table 4, at Time 1, number identification and executive function skills predicted membership in the proficient class, relative to the random class. At Time 2, number comparison and executive function skills predicted membership in the proficient class, relative to the random class. Thus, as hypothesized, children's symbolic number skills and executive function skills are important concurrent skills for number line performance in prekindergarten.

Because there were three different classes at Time 3, three sets of comparisons were made to determine whether symbolic number skills predicted class membership for (a) the proficient versus the origin class, (b) the two-endpoints versus the origin class, and (c) the proficient versus the two-endpoints class. As shown in Table 4, number identification predicted membership in the proficient class compared to the two-endpoints class. Number identification also predicted membership in the two-endpoints class compared to the origin class. However, neither number identification nor number comparison predicted membership in the proficient class compared to the two-endpoints class. Thus, children with better number identification skills had better number line performance at the end of kindergarten.

Table 4

Multinomial Logistic Regression Based on the Class Membership from the Unconditional LTA

Comparisons	*B*	*SE*	*p*	*Odds Ratio* [a]	*Confidence Interval*
Time 1: *Proficient* vs. Random					
Low-SES	-0.60	0.54	.264	0.55	[0.19, 1.57]
High-SES	0.96*	0.38	.012	2.61	[1.24, 5.52]
Verbal counting	-0.07	0.19	.707	0.93	[0.64, 1.36]
Number identification	0.98***	0.23	<.001	2.61	[1.69, 4.18]
Executive function	0.56*	0.22	.013	1.74	[1.12, 2.71]
Time 2: *Proficient* vs. Random					
Low-SES	-0.01	0.27	.976	0.99	[0.58, 1.70]
High-SES	-0.21	0.29	.477	0.81	[0.46, 1.44]
Verbal counting	-0.19	0.16	.242	0.83	[0.60, 1.14]
Number identification	0.32	0.17	.058	1.37	[0.99, 1.91]
Number comparison	0.43***	0.13	.001	1.53	[1.20, 1.96]
Executive function	0.42***	0.13	.001	1.52	[1.18, 1.97]
Time 3: *Proficient* vs. Origin					
Low-SES	0.25	0.34	.467	1.28	[0.66, 2.51]
High-SES	-0.35	0.38	.357	0.71	[0.34, 1.48]
Number identification	0.81***	0.17	<.001	2.26	[1.62, 3.15]
Number comparison	0.16	0.17	.349	1.17	[0.84, 1.62]
Time 3: *Two-endpoints* vs. Origin					
Low-SES	-0.43	0.41	.297	0.65	[0.29, 1.46]
High-SES	0.01	0.40	.978	1.01	[0.46, 2.20]
Number identification	0.74***	0.19	<.001	2.09	[1.44, 3.01]
Number comparison	-0.03	0.18	.878	0.97	[0.68, 1.39]
Time 3: *Proficient* vs. Two-endpoints					
Low-SES	0.68	0.35	.054	1.54	[0.99, 3.92]
High-SES	-0.36	0.33	.267	0.99	[0.37, 1.32]
Number identification	0.18	0.15	.226	1.03	[0.89, 1.62]
Number comparison	0.08	0.15	.605	0.48	[0.80, 1.46]

Note. $p < .05$*; $p < .001$***Continuous predictors were standardized. Dummy-coded variables were created for socioeconomic status (SES): Low-SES (1 = low, 0 = others) and High-SES (1 = high, 0 = others).

[a] Higher odd ratio values indicate that the predictor is associated with a higher probability of being in the non-referent class (italicized).

Question 3. What factors contribute to the development of number line estimation skills from pre-kindergarten to the end of kindergarten?

To determine whether symbolic number skills and executive function skills changed the probability of children transitioning from the random class at Time 1 to the proficient class at Time 2, or from the random class at Time 2 to the proficient or two-endpoints class at Time 3, we conducted two additional LTA models. In the first model, we included number identification and executive function at Time 1 as covariates, allowing them to influence the transition probabilities from the random class at Time 1 to the proficient class at Time 2. We found that number identification influenced the transition probability from the random class at Time 1 to the proficient class at Time 2 (p = .005), whereas executive function skills did not (p = .815).

In the second model, we included number identification, number comparison, and executive function skills at Time 2 as covariates, allowing them to influence the transition probabilities from the random class at Time 2 to the proficient and two-endpoints class at Time 3. We found that number identification influenced the transition probability from the random class at Time 2 to the proficient class at Time 3 (p = .025), but it was not related to the transition probability from the random class at Time 2 to the two-endpoints class at Time 3 (p = .145). Number comparison and executive function skills were not related to the transition probability from the random class at Time 2 to the proficient or two-endpoints class at Time 3 (ps > .05). Taken together, children with better number identification skills were more likely to transition from the random class to the proficient class from Time 1 to Time 2, and from Time 2 and Time 3. Executive function skills did not influence the transition at any time points.

Given that number identification was an important predictor of both current class membership and the transition from less to more proficient classes, post-hoc analyses were conducted to compare number identification performance across the classes (see Table 5). Performance on one-digit number identification trials was significantly better for the proficient than random class at both Time 1, $t(187.10) = 15.91$, $p < .001$, and Time 2, $t(331.60) = 6.14$, $p < .001$. Similarly, performance on two-digit number identification trials was significantly better for the proficient than random class at both Time 1, $t(68.88) = 6.11$, $p < .001$, and Time 2, $t(380) = 2.72$, $p = .007$. At Time 3, performance differed across the classes on both one-digit trials, $F(2, 365) = 8.45$, $p < .001$, and two-digit trials, $F(2, 359) = 16.56$, $p < .001$. More specifically, the proficient and two-endpoints classes had better performance for both one- and two-digit trials than the origin class (all *p*s < .05). However, performance on one- and two-digit number identification trials did not differ between the proficient and two-endpoints classes (all *p*s > .05).

Table 5

Mean Percentage Correct for One- and Two-Digit Number Identification Trials by Class

	Time 1		Time 2		Time 3		
Trials	Random	Proficient	Random	Proficient	Origin	Two-Endpoints	Proficient
One-Digit	52%	89%	80%	93%	91%	96%	97%
Two-Digit	11%	32%	20%	27%	38%	59%	61%

Discussion

In the present study, we examined how 4- and 5-year-old children approached a 0-10 number line task and how their number line estimation skills progressed over two years of preschool education. At the beginning of pre-kindergarten, 86% of the children had patterns of estimation that suggested they were randomly placing numbers on the number line. Cankaya et al. (2014) reported similarly poor performance for 3- to 4-year-old children. At the end of pre-kindergarten, 56% of children were still responding randomly. The rest showed evidence that they understood the task and were systematically placing target numbers on the line (Petitto, 1990; Newman & Berger, 1984; Xu & LeFevre, 2016). One year later, at the end of kindergarten, most children showed evidence that they understood the number line task. Thus, the proportion of children who could place numbers accurately on a 0-10 number line increased dramatically from the beginning of pre-kindergarten to the end of kindergarten. For some children, however, proficient performance was attained early whereas for others, performance was still weak by the end of kindergarten. Overall, the patterns of error at the beginning and end of pre-kindergarten suggest that, prior to formal schooling, many children do not have the precursor skills needed to perform the number line estimation task. As described below, concurrent number line performance was linked to executive function skills, number identification and number comparison, whereas improvements in performance were linked only to number identification skills.

Development of Number Line Estimation Skills

First, we asked how number line estimation develops prior to primary schooling (Question 1). As expected, the proportion of children who accurately placed numbers increased over time. More interestingly, development suggested that there were both quantitative and qualitative changes in this time frame. Qualitatively, at each time point, some children showed proficient performance. At all of the time points, children in the proficient class appeared to use one or more reference points to place estimates and had relatively accurate placement for numbers farther away from the endpoints, suggesting that children understood both ordinal and spatial relations among the numbers from 0 to 10 (Ashcraft & Moore, 2012; Sullivan & Barner, 2014; Barth & Paladino, 2011; Xu, 2019). Although the proportion of children in the proficient class increased over time, performance within this class was very similar across time.

The less-proficient children at Times 1 and 2 showed similarly poor performance. Moreover, although there were quantitative differences in performance between the proficient and less-proficient groups, of most interest is that number line performance varied qualitatively, with children either demonstrating an understanding of the task or not. In contrast, at Time 3 (end of kindergarten), although children in the origin and two-endpoints classes were less accurate and their estimates were more variable across targets than children in the proficient class, the difference in performance was quantitatively, rather than qualitatively, worse. That is, most children, regardless of class, demonstrated an understanding of the task and appeared to use reference points to some degree at Time 3, but children in the origin and two-endpoints classes had not yet developed the necessary skills to support accurate target placement across the whole extent of the number line. Thus, the present results add to the literature by showing that children experience both quantitative and qualitative changes in their number line performance in prekindergarten and kindergarten.

We inferred strategy use based on the patterns of error in the number line task. Children in the proficient class at all time points showed inverted U-shaped performance. Their relatively good performance at the left and right ends of the number line suggests use of reference points, a strategy that indicates an understanding of relative position, at least for the 0-10 number line. There was no evidence that children routinely used a midpoint reference, however. Children may need prompting or require additional training to use a midpoint reference at this developmental stage (Xu & LeFevre, 2016, 2018). In summary, children in the proficient class appeared to have an understanding of the number line in terms of relative position of numbers, although they probably continued to use counting on some trials, as shown by the greater error for targets farther from the endpoints.

At Time 3, children who were in the two-endpoints or origin classes also showed some evidence of using reference points, either both endpoints or primarily the origin. This conclusion is based on the finding that, for children in the two-endpoints class, the pattern of estimation errors was a pronounced inverted U shape. These children estimated numbers near the endpoints (i.e., 1, 2, 8, 9) much more accurately than numbers near the middle (i.e., 4, 5, 6). In contrast, estimation errors of children in the origin class increased steeply with each target number, such that they were much more accurate near the origin (i.e., 1, 2, 3) than near the right endpoint (i.e., 7, 8, 9). These children may have counted up from the origin, but their counting increments were not equally spaced or proportionally accurate (Xu & LeFevre, 2016). For children in the origin class, their PAE of approximately 40% for the target numbers 8 and 9 means they placed these numbers closer to the middle of the line than to the right endpoint. These results support the view that, in general, children progress from using a single reference point to using two reference points to place numbers on the number line (Petitto, 1990; Newman & Berger, 1984; Xu & LeFevre, 2016). Knowledge of reference points was not sufficient, however, to produce proficient estimation performance across the number line. Thus, although children in the origin and two-endpoints classes could identify numerical symbols, they may have lacked knowledge of the equal spacing between sequential numbers (Laski & Siegler, 2007; Xu & LeFevre, 2016).

Factors that Contribute to the Development of Number Line Estimation Skills

Second, we asked which factors concurrently relate to number line estimation for children in pre-kindergarten and kindergarten (Question 2). We found that symbolic number knowledge and executive function skills differentiated children in the random and proficient classes at Times 1 and 2. Symbolic number knowledge (i.e., number identification) was also a predictor of number line performance at Time 3, where it differentiated children in the two-endpoints class from children in the origin class and children in the proficient class from those in the two-endpoints class.

In general, research has shown that symbolic number knowledge in kindergarten is predictive of mathematics skills in early grades of elementary school (Hawes et al., 2013; Kolkman et al., 2013; LeFevre et al., 2013). For example, Göbel et al. (2014) found that children's ability to name digits prior to starting school predicted their arithmetic skills one year later, independent of their earlier arithmetic skills. Purpura et al. (2013) found that preschoolers' (3- to 5-year-old) skill at naming digits and connecting quantities to digits fully mediated the relation between their informal mathematics (i.e., counting, word problems, naming quantities) and arithmetic skills measured one year later. Thus, symbolic number knowledge is linked to the development of arithmetic skills from kindergarten to primary school.

Specific to number line estimation, we assume that if children are unable to identify numbers, they will also be unable to differentiate their locations on the number line. This assumption was supported in the present study: Children in the random class had more difficulty identifying (i.e., verbally naming) single-digit numbers at Times 1 and 2 than children in the proficient class. Furthermore, to accurately place numbers on a number line, children need an understanding of number magnitude. For example, if a child does not understand that 7 is less than 10, they will be unable to accurately place 7 on the number line. The symbolic number comparison task captures children's understanding of number magnitude (Hawes et al., 2019). Accordingly, Daker and Lyons (2018) found that symbolic number comparison predicted number line performance for first grade students. Similarly, the present research suggests that children's ability to map visual Arabic digits to verbal number labels and their knowledge of cardinal (i.e., relative magnitude) relations among the digits are necessary skills required to accurately place numbers on a number line.

Domain-general skills, specifically a composite index of executive function, differentiated random responders from children in the proficient classes at Time 1 and Time 2. Executive function skills are important for mathematics in general (see a review by Clements et al., 2016): Various executive functions allow children to monitor and manipulate information, to suppress irrelevant information, and to think flexibly when solving mathematical problems (see a review by Cragg & Gilmore, 2014). For number line estimation, children may rely on working memory, inhibition, and updating skills to maintain, manipulate, and update information relevant to their selected strategy, while inhibiting less efficient strategies and irrelevant information (Kolkman et al., 2013; Ren et al., 2019). The role of executive function in performance on number line estimation is thus consistent with previous work (e.g., Geary et al., 2008; Kolkman et al., 2013; Laski & Dulaney, 2015). Notably, however, in the present research, executive function skills did not predict the transition from the random to the proficient class. These results suggest that executive function skills are required for performing the number line task, however, the qualitative improvements from a random class to a proficient class may depend primarily on number knowledge, that is, the ability to map visual Arabic digits to verbal number labels.

Third, we asked which factors contribute to the development of number line estimation skills from pre-kindergarten to the end of kindergarten (Question 3). Although number identification, number comparison, and executive functioning were related to concurrent class membership, only number identification predicted the transition from the random class at Time 1 to the proficient class at Time 2, or from the random class at Time 2 to the proficient or the endpoints classes at Time 3. Thus, children's ability to identify numbers within the number line range (i.e., 0 to 10) was a necessary precursor for proficient number line estimation.

Although number identification predicted the transition, it did not account for all of the variance. Thus, a second possible explanation is that we did not include a key skill that might have also predicted the transition, that is, ordinal knowledge. To accurately estimate, children need to understand how numbers are related to one another. Xu and LeFevre (2016) found that 3- and 5-year-old children who were trained on sequential relations among numbers showed significant improvements in their number line estimation from pre-test to post-test. Furthermore, Xu (2019) found that children in grade 1 and 2 who had stronger understanding of ordinal associations among numbers were more likely to transition from using counting strategies to using more efficient strategies that involve estimating the location of the target number in relation to the location of multiple reference points. Thus, children in the proficient class may have had a better understanding of ordinal relations among numbers, but the mathematics measures used as predictors in this study did not tap into this knowledge. Overall, although the covariates were correlated with number line estimation performance at all three time points, the role that these variables may or may not play in the transition from less-proficient to more-proficient performance warrants further investigation.

Limitations and Future Research

In the present study, we used LCA to group children based on their patterns of estimation error. This analysis is useful because it can be used to classify children into groups that show similar performance and provide information about how estimation accuracy varies across target numbers (Oberski, 2016). However, although the classifications suggested that children used different strategies, such as counting from the origin and counting from both endpoints, nevertheless, we still relied on patterns of estimation error to infer children's strategies because we did not have a direct measure of strategy use. In contrast, Xu (2019) collected self-reports of strategy use for children in grade 1 and 2. Xu found that the self-reports corroborated the patterns reflected in the classifications. In the present research however, self-reports of strategies may not have been helpful because 4- and 5-year-olds are less able to articulate their solution processes.

Other types of analyses have been used to infer strategy use, such as model fits of congregated data from linear and logarithmic functions (e.g., Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003), or cyclical-power models (e.g., Barth & Paladino, 2011; Slusser et al., 2013). However, as in the current research (see Supplementary material) those model fits on congregated data often do not align with inferences about strategies based on other dependent measures, such as detailed error patterns for specific targets (Ashcraft & Moore, 2012; Berteletti et al., 2010), patterns of eye-tracking data (Di Lonardo et al., 2020; Di Lonardo Burr & LeFevre, 2020), or verbal strategy reports (Xu, 2019). In the future, researchers could use detailed observational reports or eye-tracking methods to more directly investigate young children's number line strategies and corroborate inferences about strategies based on patterns of performance.

Implications of the Present Research

The present research is the first longitudinal study examining the developmental trajectories of children's number line estimation from pre-kindergarten to the end of kindergarten. The results of the present research showed substantial variability and dramatic change, both qualitative and quantitative, in number line performance for children from ages four through six, supporting the view that the traditional analytical approach of averaging over the estimates in an age group is inappropriate (Bouwmeester & Verkoeijen, 2012; Xu, 2019). Many children in prekindergarten (i.e., ages 3 to 4), responded unsystematically on the number line task (i.e., the random estimators). These children also had poorer number identification skills than proficient estimators. Thus, for prekindergarten children, number identification may be a more sensitive and meaningful tool for assessing early numeracy knowledge than number line estimation. In contrast, by the end of kindergarten (i.e., ages 5 to 6), the number line estimation task may be an appropriate research tool because many children will have developed the necessary precursor symbolic number skills. Overall, the findings suggest that performance on the number line task improved dramatically as children gained knowledge of symbolic numbers.

References

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2008). Evaluating the validity of the automated working memory assessment. *Educational Psychology*, *28*(7), 725–734. https://doi.org/10.1080/01443410802243828

Ashcraft, M. H., & Moore, A. M. (2012). Cognitive processes of numerical estimation in children. *Journal of Experimental Child Psychology*, *111*(2), 246–267. https://doi.org/10.1016/j.jecp.2011.08.005

Baharudin, R., & Luster, T. (1998). Factors related to the quality of the home environment and children's achievement. *Journal of Family Issues*, *19*(4), 375–403. https://doi.org/10.1177/019251398019004002

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift: Development of numerical estimation. *Developmental Science*, *14*(1), 125–135. https://doi.org/10.1111/j.1467-7687.2010.00962.x

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological methods, 8*(3), 338. https://doi.org/10.1037/1082-989X.8.3.338

Benavides-Varela, S., Butterworth, B., Burgio, F., Arcara, G., Lucangeli, D., & Semenza, C. (2016). Numerical activities and information learned at home link to the exact numeracy skills in 5–6 years-old children. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00094

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, *46*(2), 545–551. https://doi.org/10.1037/a0017887

Blair, C., Knipe, H., & Gamson, D. (2008). Is there a role for executive functions in the development of mathematics ability? *Mind, Brain, and Education, 2*(2), 80-89. https://doi.org/10.1111/j.1751-228X.2008.00036.x

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2012). Multiple representations in number line estimation: A developmental shift or classes of representations? *Cognition and Instruction*, *30*(3), 246–260. https://doi.org/10.1080/07370008.2012.689384

Cankaya, O., LeFevre, J.-A., & Dunbar, K. (2014). The role of number naming systems and numeracy experiences in children's rote counting: Evidence from Turkish and Canadian children. *Learning and Individual Differences*, *32*, 238–245. https://doi.org/10.1016/j.lindif.2014.03.016

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, *36*, 79–90. https://doi.org/10.1016/j.ecresq.2015.12.009

Cortés Pascual, A., Moyano Muñoz, N., & Quilez Robres, A. (2019). The relationship between executive functions and academic performance in primary education: *Review and Meta-Analysis. Frontiers in psychology, 10,* 1582. https://doi.org/10.3389/fpsyg.2019.01582

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68. https://doi.org/10.1016/j.tine.2013.12.001

Daker, R. J., & Lyons, I. M. (2018). Numerical and non-numerical predictors of first graders' number-line estimation ability. *Frontiers in Psychology*, *9*, 2336. https://doi.org/10.3389/fpsyg.2018.02336

Di Lonardo, S. M., Huebner, M. G., Newman, K., & LeFevre, J.-A. (2020). Fixated in unfamiliar territory: Mapping estimates across typical and atypical number lines. *Quarterly Journal of Experimental Psychology*, *73*(2), 279–294. https://doi.org/10.1177/1747021819881631

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Di Lonardo Burr, S., & LeFevre, J. A. (2020). Fixated in more familiar territory: Providing an explicit midpoint for typical and atypical number lines. *Quarterly Journal of Experimental Psychology*, 1747021820967618.

Friso-van den Bos, I., Van Luit, J. E. H., Kroesbergen, E. H., Xenidou-Dervou, I., Van Lieshout, E. C. D. M., Van der Schoot, M., & Jonkman, L. M. (2015). Pathways of number line development in children: Predictors and risk for adverse mathematical outcome. *Zeitschrift Für Psychologie*, *223*(2), 120–128. https://doi.org/10.1027/2151-2604/a000210

Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology*, *33*(3), 277–299. https://doi.org/10.1080/87565640801982361

Geiser, C. (2013). *Data analysis with Mplus*. The Guilford Press.

Gimbert, F., Camos, V., Gentaz, E., & Mazens, K. (2019). What predicts mathematics achievement? Developmental change in 5-and 7-year-old children. *Journal of experimental child psychology, 178*, 104-120. https://doi.org/10.1016/j.jecp.2018.09.013

Göbel, S. M., Watson, S. E., Lervåg, A., & Hulme, C. (2014). Children's arithmetic development: It is number knowledge, not the approximate number sense, that counts. *Psychological Science*, *25*(3), 789–798. https://doi.org/10.1177/0956797613516471

Gravemeijer, K. (2014). Number lines in mathematics education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 466–470). Springer Netherlands. https://doi.org/10.1007/978-94-007-4978-8_121

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number line. *Developmental Psychology*, *48*(5), 1229–1241. https://doi.org/10.1037/a0027433

Hawes, Z., Nosworthy, N., Archibald, L., & Ansari, D. (2019). Kindergarten children's symbolic number comparison skills relates to 1st grade mathematics achievement: Evidence from a two-minute paper-and-pencil test. *Learning and Instruction*, *59*, 21–33. https://doi.org/10.1016/j.learninstruc.2018.09.004

Huber, S., Moeller, K., & Nuerk, H.-C. (2014). Dissociating number line estimations from underlying numerical representations. *Quarterly Journal of Experimental Psychology*, *67*(5), 991–1003. https://doi.org/10.1080/17470218.2013.838974

Hume, T., & Hume, S. (2014). *EstimationLine [Mobile app]*. https://hume.ca/ix/estimationline/

Hume, T., & Hume, S. (2014). *PathSpan [Mobile app]*. https://hume.ca/ix/pathspan/

Kolkman, M. E., Hoijtink, H. J. A., Kroesbergen, E. H., & Leseman, P. P. M. (2013). The role of executive functions in numerical magnitude skills. *Learning and Individual Differences*, *24*, 145–151. https://doi.org/10.1016/j.lindif.2013.01.004

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2014). Involvement of working memory in longitudinal development of number-magnitude skills. *Infant and Child Development, 23*, 36–50. https://doi.org/10.1002/icd

Lanza, S. T., & Cooper, B. R. (2016). Latent class analysis for developmental research. *Child Development Perspectives*, *10*(1), 59–64. https://doi.org/10.1111/cdep.12163

Laski, E. V., & Dulaney, A. (2015). When prior knowledge interferes, inhibitory control matters for learning: The case of numerical magnitude representations. *Journal of Educational Psychology*, *107*(4), 1035–1050. https://doi.org/10.1037/edu0000034

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, *78*(6), 1723–1743. https://doi.org/10.1111/j.1467-8624.2007.01087.x

Laski, E. V., & Yu, Q. (2014). Number line estimation and mental addition: Examining the potential roles of language and education. *Journal of Experimental Child Psychology*, *117*, 29–44. https://doi.org/10.1016/j.jecp.2013.08.007

LeFevre, J.-A., Cankaya, O., Xu, C., & Jimenez Lira, C. (2018). Linguistic and experiential factors as predictors of young children's early numeracy skills. In D. B. Berch, D. C. Geary, & K. Mann Koepke, *Mathematical Cognition and Learning* (pp. 49-72). Academic Press. http://dx.doi.org/10.1016/B978-0-12-812574-8.00003-1

LeFevre, J.-A., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x

LeFevre, J.-A., Jimenez Lira, C., Sowinski, C., Cankaya, O., Kamawar, D., & Skwarchuk, S.-L. (2013). Charting the role of the number line in mathematical development. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00641

Link, T., Nuerk, H.-C., & Moeller, K. (2014). On the Relation between the Mental Number Line and Arithmetic Competencies. *Quarterly Journal of Experimental Psychology*, *67*(8), 1597–1613. https://doi.org/10.1080/17470218.2014.892517

Luwel, K., Peeters, D., Dierckx, G., Sekeris, E., & Verschaffel, L. (2018). Benchmark-based strategy use in atypical number lines. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *72*(4), 253–263. https://doi.org/10.1037/cep0000153

Morgan, G. B. (2015) Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling, 22*(1), 76-86, doi: 10.1080/10705511.2014.935751

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., Pratt, M. E., Moriguchi, Y., & Loher, S. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology, 5*, 599. https://doi.org/10.3389/fpsyg.2014.00599

Muldoon, K., Simms, V., Towse, J., Menzies, V., & Guoan Yue. (2011). Cross-cultural comparisons of 5-year-olds' estimating and mathematical ability. *Journal of Cross-Cultural Psychology*, *42*(4), 669–681. https://doi.org/10.1177/0022022111406035

Muldoon, K., Towse, J., Simms, V., Perra, O., & Menzies, V. (2013). A longitudinal analysis of estimation, counting skills, and mathematical ability across the first school year. *Developmental Psychology*, *49*(2), 250–257. https://doi.org/10.1037/a0028240

Nagin, D. S. (2005). *Group-based modeling of development.* Harvard University Press. http://dx.doi.org/10.4159/9780674041318

Newman, R. S., & Berger, C. F. (1984). Children's numerical estimation: Flexibility in the use of counting. *Journal of Educational Psychology*, *76*(1), 55–64. https://doi.org/10.1037/0022-0663.76.1.55

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and nonsymbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PLoS ONE*, *8*(7), e67918. https://doi.org/10.1371/journal.pone.0067918

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science, 4*(4), 440. Doi: 10.1037/tps0000176

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling, 23*(6), 782-797. https://doi.org/10.1080/10705511.2016.1221313

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture Modeling: A monte carlo simulation study. *Structural Equation Modeling*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Nylund-Gibson, K., Grimm, R., Quirk, M., & Furlong, M. (2014). A Latent Transition Mixture Model Using the Three-Step Specification. *Structural Equation Modeling*, *21*(3), 439–454. https://doi.org/10.1080/10705511.2014.915375

Oberski, D. L. (2016). A review of latent variable modeling with R. *Journal of Educational and Behavioral Statistics*, *41*(2), 226–233. https://doi.org/10.3102/1076998615621305

Peeters, D., Degrande, T., Ebersbach, M., Verschaffel, L., & Luwel, K. (2016). Children's use of number line estimation strategies. *European Journal of Psychology of Education*, *31*(2), 117–134. https://doi.org/10.1007/s10212-015-0251-z

Peeters, D., Sekeris, E., Verschaffel, L., & Luwel, K. (2017). Evaluating the effect of labeled benchmarks on children's number line estimation performance and strategy use. *Frontiers in Psychology*, *8*, 1082. https://doi.org/10.3389/fpsyg.2017.01082

Petitto, A. L. (1990). Development of numberline and measurement concepts. *Cognition and Instruction*, *7*(1), 55–78. https://doi.org/10.1207/s1532690xci0701_3

Praet, M., & Desoete, A. (2014). Number line estimation from kindergarten to grade 2: A longitudinal study. *Learning and Instruction*, *33*, 19–28. https://doi.org/10.1016/j.learninstruc.2014.02.003

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology*, *105*(2), 453–464. https://doi.org/10.1037/a0031753

Purpura, D. J., & Ganley, C. M. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology, 122*, 104-121. https://doi.org/10.1016/j.jecp.2013.12.009

Ramani, G. B., Jaeggi, S. M., Daubert, E. N., & Buschkuehl, M. (2017). Domain-specific and domain-general training to improve kindergarten children's mathematics. *Journal of Numerical Cognition*, *3*(2), 468–495. https://doi.org/10.5964/jnc.v3i2.31

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, *79*(2), 375–394. https://doi.org/10.1111/j.1467-8624.2007.01131.x

Ren, K., Lin, Y., & Gunderson, E. A. (2019). The role of inhibitory control in strategy change: The case of linear measurement. *Developmental psychology*, *55*(7), 1389. https://doi.org/10.1037/dev0000739

Sasanguie, D., Van den Bussche, E., & Reynvoet, B. (2012). Predictors for mathematics achievement? Evidence from a longitudinal study. *Mind, Brain, and Education*, *6*(3), 119–128. https://doi.org/10.1111/j.1751-228X.2012.01147.x

Schneider, M., Merz, S., Stricker, J., Smedt, B. D., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, *89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

Siegler, R. S., & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science*, *19*(8), 759–763. https://doi.org/10.1111/j.1467-9280.2008.02153.x

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–250. https://doi.org/10.1111/1467-9280.02438

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, *11*(5), 655–661. https://doi.org/10.1111/j.1467-7687.2008.00714.x

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, *101*(3), 545–560. https://doi.org/10.1037/a0014239

Slusser, E. B., & Barth, H. C. (2017). Intuitive proportion judgment in number-line estimation: Converging evidence from multiple tasks. *Journal of Experimental Child Psychology*, *162*, 181–198. https://doi.org/10.1016/j.jecp.2017.04.010

Sullivan, J., & Barner, D. (2014). The development of structural analogy in number-line estimation. *Journal of Experimental Child Psychology*, *128*, 171-189. https://doi.org/10.1016/j.jecp.2014.07.004

Sutherland, M., Clarke, B., Nese, J. F., Cary, M. S., Shanley, L., Furjanic, D., & Durán, L. Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery. *Early Childhood Research Quarterly*, *55*, 119-128. https://doi.org/10.1016/j.ecresq.2020.11.003

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768–1786. https://doi.org/10.1111/j.1467-8624.2010.01509.x

Whyte, J. C., & Bull, R. (2008). Number games, magnitude representation, and basic number skills in preschoolers. *Developmental Psychology*, *44*(2), 588–596. https://doi.org/10.1037/0012-1649.44.2.588

Xu, C. (2019). Ordinal skills influence the transition in number line strategies for children in Grades 1 and 2. *Journal of Experimental Child Psychology*, *185*, 109–127. https://doi.org/10.1016/j.jecp.2019.04.020

Xu, C., & LeFevre, J.-A. (2016). Training young children on sequential relations among numbers and spatial decomposition: Differential transfer to number line and mental transformation tasks. *Developmental Psychology*, *52*(6), 854–866. https://doi.org/10.1037/dev0000124

Xu, C., & LeFevre, J.-A. (2018). Cross-cultural comparisons of young children's early numeracy performance: Effect of an explicit midpoint on number line performance for Canadian and Chinese-Canadian children, *Bordón, 70*, 123-138. https://doi.org/10.13042/Bordon.2018.60966

Xu, X., Chen, C., Pan, M., & Li, N. (2013). Development of numerical estimation in Chinese preschool children. *Journal of Experimental Child Psychology*, *116*(2), 351–366. https://doi.org/10.1016/j.jecp.2013.06.009

Yuan, L., Prather, R., Mix, K. S., & Smith, L. B. (2019). Number representations drive number-line estimates. *Child Development*, cdev.13333. https://doi.org/10.1111/cdev.13333

Zippert, E. L., & Ramani, G. B. (2017). Parents' estimations of preschoolers' number skills relate to at-home number-related activity engagement: Parents' estimates of child's number skills. *Infant and Child Development*, *26*(2), e1968. https://doi.org/10.1002/icd.1968