Journal of Numerical Cognition
jnc.psychopen.eu | 2363-8761

Research Reports

A Comparison of Methods for Assessing Performance on the Number Line Estimation Task

Susan I. Gross[a], Carol A. Gross[a], Dan Kim[b], Sarah L. Lukowski[b], Lee A. Thompson*[a], Stephen A. Petrill[b]

[a] Department of Psychological Sciences, Case Western Reserve University, Cleveland, OH, USA. **[b]** Department of Psychology, The Ohio State University, Columbus, OH, USA.

Abstract

The debate about how to characterize performance on the number line estimation (NLE) task has yielded a diverse set of accuracy measures. These accuracy measures include characterizing performance by deviation from the correct score with percent absolute error (PAE), modeling the shape of responses via the logarithmic-to-linear shift, and modeling the strategy use via the cyclical power model (one and two cycle). In the present study, accuracy on a symbolic NLE task was examined using phenotypic and quantitative genetic analyses of all four measurements. Data were collected from a same-sex twin sample at ages 12 and 15 (N = 150 pairs) as part of the Western Reserve Reading and Math Project. Linear mixed-effect models were used to compare how well the four NLE accuracy measures predicted math achievement, as measured by the Woodcock Johnson-III Fluency, Calculation, and Applied Problems subtests, after cognitive ability was controlled. NLE accuracy measures were not related to Fluency or Calculation after cognitive ability was controlled, but all NLE accuracy measures were related to Applied Problems at 12 and 15 years old. Although theories about what the NLE task measures have been contested in the literature, the relationship between NLE accuracy and achievement did not differ regardless of the type of accuracy measure used. In addition, the estimates for genetic and environmental influences were proportionately similar across the NLE accuracy measures. Overall, all proposed measures of accuracy in the present sample appear appropriate for prediction of math achievement in adolescents.

Keywords: math cognition, number line, behavior genetics, math achievement

Journal of Numerical Cognition, 2018, Vol. 4(3), 554–571, doi:10.5964/jnc.v4i3.120

Received: 2017-02-28. Accepted: 2018-02-01. Published (VoR): 2018-12-21.

*Corresponding author at: Department of Psychological Sciences, Case Western Reserve University, 10900 Euclid Ave. Cleveland, OH, 44106, USA. E-mail: lat@case.edu

In an era of increasing demand on schools to prepare students for success in STEM careers, the urgency for gaining a better understanding of the underlying components of math cognition that predict achievement in math is increasing. Pervasive core processes for mathematical reasoning include the manner and precision with which people conceptualize, represent, manipulate, and compare internal numerical magnitude representation (Siegler & Lortie-Forgues, 2014) and involves both symbolic and nonsymbolic numeric representations. Fazio, Bailey, Thompson, and Siegler (2014) provide compelling evidence that symbolic and nonsymbolic numerical magnitude estimation independently predict mathematics achievement; though, the relationship between nonsymbolic numerical magnitude estimation and mathematical achievement diminishes after the first grade. However, both practical and theoretical questions about the nature of numeric magnitude representation

remain. At the present time, two theoretical approaches attempt to explain patterns of performance on numeric estimation tasks. The first emphasizes the importance of an innate internal mental representation of magnitude which takes the form of a number line (Dehaene, 1992). A second perspective suggests that the development and application of strategies determines performance (Barth & Paladino, 2011). Depending on the theoretical orientation, the recommended approach for quantifying numeric estimation accuracy is different, thus leading to a separate but related practical question in regard to how performance on numeric estimation tasks is best represented. For example, the predicted shape of response is different depending on whether performance on the task is driven by an innate magnitude representation (predicts shift from logarithmic to linear response) or by strategy (predicts minimal error around the endpoints and midpoint). The practical disagreement about which measure to use is relevant because higher performance on the numeric magnitude estimation tasks has been associated with higher levels of mathematical achievement in elementary school children when average error was used (Ashcraft & Moore, 2012; Booth & Siegler, 2006; Fazio et al., 2014; Geary, 2011; Sasanguie, De Smedt, Defever, & Reynvoet, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013; Siegler & Booth, 2004). The alternate, theory-driven, measures have been used to characterize performance, but a comparison of the predictive validity of the accuracy measures on math achievement outcomes in an adolescent sample has not been conducted.

More broadly, when symbolic numbers (e.g., “8”) were presented to both adults and children, regardless of modality, magnitude representation was automatically activated, even in cases where a task did not require a judgment of magnitude (Wood, Willmes, Nuerk, & Fischer, 2008). Internal magnitude representation is ordinal, with numbers cognitively organized so that smaller numbers are associated with the left side and larger numbers are associated with the right side in external space (Dehaene, Bossini, & Giraux, 1993). Therefore, subjects respond more quickly to larger numbers with their right hands and smaller numbers with their left hands. In addition, subjects respond more quickly to comparisons when numbers are farther apart from one another than when they are closer together (distance effect), indicating that people have stronger connections between numbers that are closer on a number line (Duncan & McFarland, 1980). Together, this research has provided evidence for internal magnitude representation as a “mental number line.”

Internal magnitude representation has been measured using a variety of techniques (Barth & Paladino, 2011; Bouwmeester & Verkoeijen, 2012; Cohen & Blanc-Goldhammer, 2011; Peeters, Degrande, Ebersbach, Verschaffel, & Luwel, 2016; Sasanguie & Reynvoet, 2013; Slusser, Santiago, & Barth, 2013). The most popular task used to measure how accurately subjects relate number to space concretely is the number line estimation (NLE) task (Siegler & Opfer, 2003). The NLE task requires subjects to either mark on a number line where a given number should be placed (number to position NLE task) or identify the number that corresponds to a spot already marked on a number line (position to number NLE task). The present study and review focus on tasks that use the number to position NLE task with symbolic endpoints (Arabic numbers), although there is also research that considers effects when the NLE task has non-symbolic endpoints such as dots.

Although variants of the procedures exist, the symbolic NLE task is arguably the most well-studied of the variants, and characteristics of the symbolic NLE task have been studied in relation to performance on other tasks as well. Over time, students become better on the NLE task, and individual differences in performance on the NLE task are associated with higher math achievement on curriculum based tests in elementary school (Booth & Siegler, 2006; Sasanguie et al., 2012; Sasanguie et al., 2013; Siegler & Booth, 2004). The relationship is also significant after controlling for IQ, indicating that performance on the NLE task is specifically predictive of math-

ematics skill acquisition beyond general intelligence (Geary, 2011). Performance on the NLE task may not predict all aspects of math achievement, however; a correlation between the average error on the NLE task and accuracy on a timed arithmetic test for first through third grade subjects was not significant, but the correlation between the average error on the NLE task and the general curriculum-based math achievement task was significant (Sasanguie et al., 2013). Therefore, there may be specific aspects of the NLE task that are related to performance on more general curriculum-based tests that are not found in timed arithmetic fact tests.

As with math achievement, there are several ways to evaluate performance on the NLE task, and the most common for the NLE task is average error. In addition to the average error, the pattern of response has also been identified as a relevant aspect of the task, with some children displaying a logarithmic response pattern and others displaying a linear response pattern (Booth & Siegler, 2006; Siegler & Booth, 2004; Siegler & Opfer, 2003; Siegler, Thompson, & Opfer, 2009). Logarithmic responding is characterized by a pattern of response in which the subject affords more space to smaller numbers on the left side of the number line and then condenses the spaces between larger numbers on the right side of the number line. Linear responding is a pattern of accurate response in which the child equally spaces numbers from left to right. The majority of kindergarteners respond logarithmically to a 0-100 number line, but the majority of second grade students respond linearly (Booth & Siegler, 2006; Siegler & Booth, 2004). Response patterns appear to undergo a logarithmic-to-linear shift as particular numbers become familiar, so children can simultaneously have a linear representation in one order of magnitude (e.g., 0-100 number line) and a logarithmic representation in another order of magnitude (e.g., 0-1000 number line); this is the case for the majority of second grade students, who are less familiar with larger magnitudes (Booth & Siegler, 2006; Siegler & Opfer, 2003). By sixth grade, students are more likely to respond in a linear fashion to both the 0-100 and 0-1000 number lines (Booth & Siegler, 2006; Siegler & Opfer, 2003).

With the ability to hold multiple representations at the same time, it appears that the logarithmic-to-linear shift in performance on the NLE task may be driven by a familiarity with numbers that arises from school instruction. Further evidence of the importance of school instruction comes from the fact that adults from cultures without formal mathematical instruction map number to space logarithmically, whereas the majority of adults from Western cultures map number to space linearly (Dehaene, Izard, Spelke, & Pica, 2008). While the logarithmic-to-linear shift appears to be a product of formal education, it does not occur at the same time for all children. Children with dyscalculia, for example, tend to show more logarithmic response patterns than children with typical development from ages 8 to 10 (Kucian et al., 2011).

Due to the developmental logarithmic-to-linear shift, linearity has been used as a performance measure on the NLE task. Mathematics achievement has been correlated with linearity based on individual fit statistics to linear functions using R^2 (Ashcraft & Moore, 2012; Booth & Siegler, 2006; Siegler & Booth, 2004). However, R^2 reflects the degree of fit to a linear function but does not account for the degree of linearity compared to logarithmicity. Therefore, a function may be low in R^2 due to deviation from the correct answer on a set of items, but the pattern of response may still more appropriately fit a linear function than a logarithmic function. A mixed log-linear model (MLLM) may be more appropriate for specifically identifying the degree of logarithmic responding (Anobile, Cicchini, & Burr, 2012; Cicchini, Anobile, & Burr, 2014). The degree of logarithmic responding in the MLLM fit for performance of kindergarteners (0-30), first graders (0-100), and second graders (0-1000) was found to be significantly correlated with accuracy on addition and subtraction problems (Kim & Opfer, 2017).

Although there is robust evidence to suggest that performance on the NLE task is better in older and more well-educated children, the connection between better performance on the NLE task and changes to internal magnitude representation is still under debate (Barth & Paladino, 2011; Bouwmeester & Verkoeijen, 2012; Cohen & Blanc-Goldhammer, 2011; Peeters et al., 2016; Sasanguie & Reynvoet, 2013; Slusser et al., 2013). The lower errors and increased linearity on the NLE task may be due to a change in the subjects' strategy use and familiarity with the subject matter rather than due to a change in the subject's internal magnitude representation. The influence of strategy in the NLE task is demonstrated through the pattern of lower errors around important markers, specifically the endpoints and midpoint (Bouwmeester & Verkoeijen, 2012). If a true change in internal magnitude representation was occurring, lower errors around midpoints and endpoints would not be predicted because internal magnitude representation is continuous, unbounded, and does not have an identifiable midpoint (Bouwmeester & Verkoeijen, 2012). When the midpoint is marked on the NLE task, accuracy is even higher at the midpoint (Peeters et al., 2016). In addition, subjects who talk more about using markers as guides for their responses perform better on the task and have higher mathematics achievement (Peeters et al., 2016).

Not only does attention to the midpoint and endpoints change performance, but also the range of numbers used for the number line changes the pattern of responding (Hurst, Leigh Monahan, Heller, & Cordes, 2014). College-aged students who responded linearly on number lines with familiar endpoints responded logarithmically on number lines with non-familiar endpoints and also responded logarithmically to a number line with letter anchors (Hurst et al., 2014). Thus, the logarithmic-to-linear shift witnessed in early elementary school for the majority of students may be due to strategy use and familiarity with the stimuli rather than a change in internal magnitude representation.

If responses on the NLE task are driven by strategy use, specifically how subjects use the midpoint and endpoints, then performance on the task may be more accurately characterized by a function that predicts minimal errors around the markers in use, the cyclical power model (Barth & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011; Slusser et al., 2013). The cyclical power model, explained in more detail below, has directly challenged the logarithmic-to-linear shift by viewing the change in performance on the NLE task as one that evolves continuously through the use of various midpoint strategies rather than categorically, from a logarithmic-to-linear response (Barth & Paladino, 2011). In addition, the use of the cyclical power model acknowledges that the NLE task is one that requires a proportional judgment (e.g., given the total length of the number line, how much space should I allot to this given number?). True internal magnitude representation is not bounded and thus not one of proportion judgments, so the use of cyclical power model is intended not to more closely resemble changes in internal magnitude representation but to instead capture strategy use by the participant. The specific strategy use that is predicted determines the type of cyclical power model used, with a one-cycle cyclical power model appropriate for participants using the endpoints as markers, and the two-cycle cyclical power model appropriate for participants using the endpoints and midpoints as markers.

Due to these theoretical disagreements, comparisons between the cyclical power model and the logarithmic-to-linear shift have been conducted (mostly with group level data) with mixed results. In these studies, the fit of a mixed logarithmic-linear model (MLLM) representing the logarithmic-to-linear shift is favored in some (Dackermann, Huber, Bahnmueller, Nuerk, & Moeller, 2015; Kim & Opfer, 2017; Opfer, Thompson, & Kim, 2016), while the fit of the cyclical power model is favored in others (Barth & Paladino, 2011). Although group level data has been used to compare the logarithmic-to-linear shift and the cyclical power model, the relationship between individual differences in performance based on the cyclical power model and math achievement

has not been as well established. An individual differences analysis using the model fit statistic Akaike Information Criterion (AIC) to compare the preferred fit of various models on the NLE task did show that performance on the NLE task in higher grades tended to be more appropriately captured by the cyclical power model than a model capturing the logarithmic-to-linear shift, indicating that strategy may be fundamental to task performance in higher grades (Sasanguie, Verschaffel, Reynvoet, & Luwel, 2016). In a separate study, the cyclical power model was not associated with accuracy on addition nor subtraction problems in younger subjects, but the fit for the MLLM, representing the logarithmic-to-linear shift, did significantly predict performance (Kim & Opfer, 2017).

Given the theoretical disagreement in the literature, it is not clear how well each method of scoring can predict performance on math achievement tests. The purpose of this study is to examine how different ways of measuring accuracy on the number line task using methods from theoretically different origins predict math achievement.

Quantitative Genetics

Behavior genetics is a tool used to investigate the origins of individual differences in traits by calculating the proportion of the variation accounted for by genetics, shared environment, and nonshared environment or error in a trait. Behavior genetics has particular utility in the present study because a direct measure of the amount of variation accounted for by genetics and different aspects of the environment on the NLE task has not been conducted. In addition, examining the NLE data in a behavior genetics framework will allow us to validate the theoretical underpinnings of each approach. Theoretically, the shared environmental factor, which would account for shared experiences in school settings, would be significant for the logarithmic-to-linear shift because schooling is linked to more linear responding on the task. Training programs have also shown success in initiating the logarithmic-to-linear shift (Kucian et al., 2011; Opfer & Siegler, 2007; Ramani & Siegler, 2008; Siegler & Ramani, 2009; Thompson & Opfer, 2008, 2016). Therefore, there is evidence that the shared environment such as attending the same school may affect the shape of responding on the NLE task. Schooling may also affect the individual's application of strategy, with older children being more likely to display patterns of responding consistent with the cyclical power model (Sasanguie et al., 2016). Genetics are expected to drive variation for all measures of the NLE task due to the cognitive processes required by the task as extant behavioral genetic research provides ample evidence for the pervasive influence of genes on individual differences in general cognitive ability and in specific cognitive abilities (Plomin, DeFries, Knopik, & Neiderheiser, 2013). Behavioral genetic models have not previously been applied in the same sample of research participants on NLE measures reflecting both cyclical power model and logarithmic-to-linear shift approaches; therefore, we view these analyses as exploratory. Nevertheless, if the NLE measures representing the cyclical power model and the logarithmic-to-linear shift differentially predict math achievement, behavioral genetic analyses may provide important information on whether genetic, shared, and/or nonshared environmental influences mediate the differential prediction.

Present Study

This study is unique in using an adolescent sample to evaluate individual differences in responses on a NLE task, which not only will fill a grade and age hole in the literature but will also describe individual differences at an older age range in which linear responding may be assumed but not proven for all participants. Although the logarithmic-to-linear shift for numbers 0-1000 occurs in elementary school for most children, performance still

varies in sixth grade. For example, at age 12, a logarithmic function was still the best fit function for 28% of participants (Siegler & Opfer, 2003). As further evidence that age is not a substitute for the developmental stage a child is in, some older children still display overestimation of smaller numbers while some younger children display linear responding (Bouwmeester & Verkoeijen, 2012). In fifth grade, modeling the use of the midpoint is still not universal, as only 58% of the subjects were best fit by that model (Rouder & Geary, 2014). Thus, although the transition to more sophisticated responding on this task is often studied in early to late elementary school, there is still variation in response patterns after elementary school. In addition, the present study benefits from a longitudinal design, in which stability of performance between time points can be assessed. Correlations between 5-year old performance on the NLE task at 10-week interval measurements ranged from .43 to .50 for endpoints 1-10 and .37 to .56 for endpoints 1-100 (Muldoon, Towse, Simms, Perra, & Menzies, 2013); however, a longitudinal assessment of the stability of number line estimation performance in adolescents has not been conducted.

Overall, each measurement style is attempting to measure different characteristics of performance on the NLE task. However, given that we are attempting to use the number line estimation task to understand individual differences and the predictive validity for achievement, a comparison of the accuracy measures, though they are theoretically different, is called for. The present study attempts to answer the following three questions. First, do the measures closely resemble one another? It was hypothesized that, given the theoretical differences between the measures, these accuracy measures would not be highly correlated. Second, are the accuracy measures distinct in their predictive value of different types of math achievement? It was hypothesized that NLE task performance would be most highly predictive of math achievement measures that involve complex math reasoning involving proportion judgment. Third, are the accuracy measures distinct in their genetic and environmental origins? It was hypothesized that variation in all measures would be significantly predicted by a genetic component due to the cognitive nature of the task. In addition, it was hypothesized that the shared environment component would be relevant for all measures due to past studies that have demonstrated the influences of schooling and intervention.

Method

Participants

Data were drawn from the Western Reserve Reading and Math Project (WRRMP), a 10-wave longitudinal twin study in which same-sex twin pairs were recruited from school nominations and birth records in kindergarten or first grade. Data for the present study were drawn from the 8th and 9th measurement occasions. These waves of measurement were approximately 3.0 years apart (on average), and the participants averaged 12.2 years (*SD* = 1.2) in the 8th wave and 15.4 years (*SD* = 1.4) in the 9th wave. Participants who completed the number line task at age 12 and again at age 15 and who had scores on all four accuracy measures at each time point were included for the present study (*N* = 300; MZ: *n* = 130; DZ: *n* = 170). Participants were mostly White (94%), and 56% were female. DNA genotyping was used to determine zygosity, and in cases without genotyping consent, zygosity was established using a parental questionnaire (Goldsmith, 1991).

Measures

Number Line Estimation Task

The NLE task was administered to participants at ages 12 and 15 via a pencil and paper format with 0 and 1000 displayed at opposite ends of the number line (Opfer & Siegler, 2007). The participant was instructed to identify the location of 500 on the number line before beginning the trials, and the administrator corrected mistakes made. During the trial phase, participants were presented with a new number line anchored at the ends with 0 and 1000 and a number at the top of the page and were asked to mark on the line the appropriate location of the number. In total, 22 trials were presented sequentially in the same ascending, non-randomized order for all participants (i.e., 2, 5, 18, 34, 56, 78, 100, 122, 147, 150, 163, 179, 246, 366, 486, 606, 722, 725, 738, 754, 818, 938). Accuracy on the NLE task was characterized by four measures (explained in more detail below): percent absolute error, mixed log-linear model, one-cycle cyclical power model, and two-cycle cyclical power model.

Differences in the administration procedure (emphasizing and correcting the half mark vs. not drawing attention to the half mark) and item distribution (oversampling the left side of the distribution vs. evenly sampling the distribution) has been shown to change which model (a mixed log-linear model or a mixed cyclical power model) has a better fit to the responses (Opfer et al., 2016). In the comparison of the conditions, 58.33% of the individuals in the condition consistent with the data collection procedure in the present study had response patterns that were better fit by the mixed log-linear model rather than the mixed cyclical power model. Given that the approach had almost equal best fit results for both models, the results of the present study should not be influenced heavily by the appropriateness of fit of one function over the other due to the administration procedures and sampling.

Percent absolute error — Percent absolute error (PAE) is the sum of the absolute value of errors divided by the total length of the number line times the number of trials (Equation 1).

$$\mathrm{PAE} = \sum \frac{|Number\ given - Numerical\ response|}{1000 * Number\ of\ trials}$$

Equation 1. PAE.

Mixed Log-Linear Model — The mixed log-linear model characterizes the shape of response between a linear and logarithmic function (Anobile et al., 2012; Cicchini et al., 2014). In Equation 2, *R* is a vector of the student's responses, *N* is a vector of the numbers given, *a* is a scaling parameter, and λ is the degree of logarithmic trend in the response pattern. The accuracy value targeted here is λ, which ranges from 0 (perfectly linear responding) to 1 (perfectly logarithmic responding).

$$R = a\left((1-\lambda)N + \lambda\frac{1000}{\ln(1000)}\ln(N)\right)$$

Equation 2. λ.

One-Cycle Cyclical Power Model — The one-cycle cyclical power model is a function that predicts the use of 0 and 1000 as anchors that assist in the proportion judgment (Barth & Paladino, 2011; Hollands & Dyre, 2000;

Spence, 1990). The model predicts that subjects using the endpoints as anchors will have a predictable pattern of response, overestimating on one side of the midpoint and underestimating on the other side of the midpoint.

N in this equation was divided by the total length of the number line so that the fitted values represented the trend of a proportion judgment. The data were fit with one free parameter (β_1) to indicate the shape of the function, with a value of 1 indicating a perfectly linear response, and values deviating further from 1 indicating greater distance from the linear response. All values of β_1, shown in Equation 3, were recalculated to the absolute value of the difference between their original value and 1 so that values of 0 would represent perfectly linear responding, and values further from 0 would represent less linear responding.

$$R = 1000 * \frac{\left(\frac{N}{1000}\right)^{\beta_1}}{\left(\frac{N}{1000}\right)^{\beta_1} + \left(1 - \frac{N}{1000}\right)^{\beta_1}}$$

Equation 3. β_1.

Two-Cycle Cyclical Power Model — Maturation of numerical ability would theoretically draw attention not only to the endpoints of the number line but also to its midpoint; subjects begin to use the midpoint as an anchor in the task, noting that values greater than 500 should be placed to the right of the midpoint, and values less than 500 should be placed to the left of the midpoint. The use of a midpoint strategy would lead to responses consistent with a two-cycle cyclical power model (Hollands & Dyre, 2000). The use of the midpoint creates a pattern of response in which numbers are overestimated between 0 and 250, underestimated between 250 and 500, overestimated between 500 and 750, and underestimated between 750 and 1000 as the participant judges the distance from the anchor point of their choosing or shows the opposite pattern of underestimate-overestimate-underestimate-overestimate depending on which anchors they reference. The free parameter for this model, represented in Equation 4a and 4b, is β_2. Values at 1.0 represent linear responding, and values further away from 1.0 represent greater deviations from a linear response. All values of β_2 were recalculated to the absolute value of the difference between their original value and 1 so that values of 0 would represent perfectly linear responding, and values further from 0 would represent less linear responding.

$$R = 500 * \frac{\left(\frac{N}{1000}\right)^{\beta_2}}{\left(\frac{N}{1000}\right)^{\beta_2} + \left(0.5 - \frac{N}{1000}\right)^{\beta_2}}$$

Equation 4a. β_2.

Note. $N < 500$.

$$R = 500 * \frac{\left(\frac{N}{1000} - 0.5\right)^{\beta_2}}{\left(\frac{N}{1000} - 0.5\right)^{\beta_2} + \left(1 - \frac{N}{1000}\right)^{\beta_2}} + 500$$

Equation 4b. β_2.

Note. $N \geq 500$.

Math Achievement

The Woodcock Johnson III (WJ-III) was administered as a measure of math achievement and included mathematics subtests Math Fluency, Calculation, and Applied Problems at ages 12 and 15 (Woodcock, McGrew, & Mather, 2001). Math Fluency is a measure of how quickly participants can solve simple math problems; participants were given three minutes to solve 160 addition, subtraction, and multiplication problems using paper and pencil. Performance was based on the number of problems solved correctly in the limited time. The Calculation subtest measures the ability to solve math problems using paper and pencil, ranging in difficulty from single digit addition and subtraction to geometry, trigonometry, and calculus problems. There was no time constraint on the Calculation subtest. The Applied Problems subtest measures the participant's ability to solve story problems in a free response format without a time constraint. The story problems were presented visually and read aloud to the subjects. The subtest required participants to understand the story problem and apply the appropriate mathematical procedure without assistance. Some items included spurious information that was to be ignored for successful completion of the problem. Problem difficulty ranged from basic arithmetic to more advanced topics in probability, and several problems tested proportional reasoning.

General Cognitive Ability

A general cognitive ability summary measure was compiled from several measures, including the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983), Clinical Evaluation of Language Fundamentals (CELF) Word Classes subtest (Semel, Wiig, & Second, 2003), Woodcock Reading Mastery Test — Revised (WRMT-R) Word Identification subtest (Woodcock, 1998), Wechsler Intelligence Scale for Children (WISC) Symbol Search subtest (Wechsler, 2004), and Comprehensive Test of Phonological Processing (CTOPP) Rapid Digit Naming and Rapid Letter Naming subtests (Wagner, Torgesen, & Rashotte, 1999) administered at age 12. The first, unrotated principal component was used as a summary measure for general cognitive ability, as was done in a previous publication (Lukowski et al., 2014).

Results

Descriptive Statistics

Descriptive statistics for raw values are listed in Table 1. Participants with values on any one NLE accuracy measure greater than 3 standard deviations from the sample mean (all on the low end of performance) were removed prior to analysis (age 12: 2; age 15: 7). PAE and λ were log transformed to correct for skewness [skew before correction (*SE*): age 12, PAE = 3.4 (.14), λ = 2.1 (.14); age 15, PAE = 1.8 (.14), λ = 3.0 (.14)]. For the remaining analyses, all variables were residualized on age and sex and standardized.

Phenotypic correlations are displayed in Table 2. All of the number line measures are arranged so that smaller values indicate more accurate responding; therefore, values with achievement measures are, as predicted, negatively correlated with number line accuracy measures. Overall, the 12-year-old NLE measures shared 91% variance, and the 15-year-old NLE measures shared 79% variance. The correlations between the time points were also significant for all of the measures (PAE: .37, λ: .44, β_1: .45, β_2: .28). Within time points, all measures of NLE predicted all three measures of math achievement except for β_2 and Fluency at 12 years old. In addition, the relationship between all 12-year-old NLE measures and 15-year-old math achievement measures were significant except for β_2 and Fluency.

Journal of Numerical Cognition
2018, Vol. 4(3), 554–571
doi:10.5964/jnc.v4i3.120

Table 1

Descriptive Statistics of Performance on the NLE Task and Mathematical Achievement Measures for 12 and 15-Years-Old

	12-years-old				15-years-old			
Task	***M* (*SD*)**	**Min**	**Max**	***N***	***M* (*SD*)**	**Min**	**Max**	***N***
NLE accuracy								
PAE (log)	3.97 (0.58)	2.79	5.99	300	3.55 (0.46)	2.28	4.85	300
λ (log)	0.12 (0.14)	0	0.66	300	0.03 (0.07)	0	0.37	300
$\lvert\beta_1\text{-}1\rvert$	0.2 (0.19)	0	0.80	300	0.08 (0.11)	0	0.54	300
$\lvert\beta_2\text{-}1\rvert$	0.35 (0.24)	0	0.90	300	0.19 (0.16)	0	0.73	300
WJ Fluency	102.07 (16.41)	68	181	292	104.59 (18.03)	63	171	300
WJ Calculation	103.65 (13.85)	52	146	287	102.9 (15.91)	62	140	299
WJ Applied Problems	108.55 (10.57)	69	134	299	107.26 (10.68)	76	140	295

Note. PAE (log) is the log-transformed percent absolute error. λ (log) is the log-transformed logarithmicity from the mixed log-linear model. $\lvert\beta_1\text{-}1\rvert$ is the corrected free parameter of the one-cycle cyclical power model. $\lvert\beta_2\text{-}1\rvert$ is the corrected free parameter of the two-cycle cyclical power model. WJ Fluency is the Woodcock-Johnson III Fluency subtest. WJ Calculation is the Woodcock-Johnson III Calculation subtest. WJ AP is the Woodcock-Johnson III Applied Problems subtest.

Table 2

Correlations Between NLE Task Accuracy Measures and Math Achievement Measures at Ages 12 and 15

Task	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
12-years-old															
1. PAE (log)		.87**	.88**	.86**	-.16**	-.25**	-.40**	.37**	.40**	.41**	.28**	-.14**	-.28**	-.36**	-.28**
2. λ (log)			.97**	.87**	-.13**	-.20**	-.37**	.34**	.44**	.44**	.30**	-.17**	-.23**	-.36**	-.25**
3. $\lvert\beta_1\text{-}1\rvert$				.88**	-.13**	-.24**	-.38**	.37**	.42**	.45**	.31**	-.17**	-.25**	-.36**	-.26**
4. $\lvert\beta_2\text{-}1\rvert$					-.07	-.21**	-.37**	.35**	.37**	.39**	.28**	-.08	-.24**	-.33**	-.26**
5. WJ Fluency						.58**	.47**	-.17**	-.14*	-.12*	-.10	.81**	.45**	.44**	.43**
6. WJ Calculation							.66**	-.31**	-.22**	-.24**	-.23**	.53**	.66**	.69**	.49**
7. WJ AP								-.43**	-.37**	-.36**	-.38**	.51**	.66**	.82**	.59**
15-years-old															
8. PAE (log)									.63**	.64**	.71**	-.22**	-.29**	-.43**	-.28**
9. λ (log)										.95**	.79**	-.17**	-.20**	-.29**	-.17**
10. $\lvert\beta_1\text{-}1\rvert$											.72**	-.15**	-.21**	-.30**	-.16**
11. $\lvert\beta_2\text{-}1\rvert$												-.14*	-.22**	-.32**	-.17**
12. WJ Fluency													.53**	.54**	.48**
13. WJ Calculation														.76**	.48**
14. WJ AP															.59**
15. g composite															

Note. PAE (log) is the log-transformed percent absolute error. λ (log) is the log-transformed logarithmicity from the mixed log-linear model. $\lvert\beta_1\text{-}1\rvert$ is the corrected free parameter of the one-cycle cyclical power model. $\lvert\beta_2\text{-}1\rvert$ is the corrected free parameter of the two-cycle cyclical power model. WJ Fluency is the Woodcock-Johnson III Fluency subtest. WJ Calculation is the Woodcock-Johnson III Calculation subtest. WJ AP is the Woodcock-Johnson III Applied Problems subtest.
*p < .05. **p < .01.

Linear Mixed-Effect Models

In order to account for non-independence of observations (due to the fact that the participants were each part of a twin dyad), linear mixed-effect models with random intercept and Satterwhaite correction were used (Kenny, Kashy, Cook, & Simpson, 2006). The lme4 package was used to perform the analysis in R (Bates, Mächler, Bolker, & Walker, 2015).

$$Y_{ij} = a_0 + d_i + b_1(\text{g factor})_{ij} + b_2(\text{NLE measure})_{ij} + e_{ij}$$

Equation 5. Linear mixed-effect model predicting achievement from NLE task performance and g.

The NLE measure is used to predict math achievement while controlling for g in Equation 5. In the equation, i represents the individual, and j represents the grouping variable (dyad). *Y* is math achievement, a_0 is the grand mean, d_i is the random intercept, b_1 is the main effect of the g factor, b_2 is the main effect of the NLE measure, and e_{ij} is error. The main effects of the NLE measure for the models are listed in Table 3. The main effect of the g factor was significant for all comparisons. The main effect of the NLE task performance after controlling for g was significant for all NLE measures at 12 and 15 years old for Applied Problems but not for Fluency and Calculation.

Table 3

Coefficients of the Linear Mixed-Effect Models Predicting Achievement From NLE Task Accuracy Measures

	12-years-old				15-years-old			
NLE	df_1	b_1	df_2	b_2	df_1	b_1	df_2	b_2
Fluency								
PAE (log)	270.74	.29***	221.98	.01	275.79	.37***	213.40	.03
λ (log)	269.96	.28***	215.78	-.01	276.91	.36***	224.01	-.04
$\lvert\beta_1\text{-}1\rvert$	269.88	.28***	216.31	-.03	276.89	.36***	222.40	-.01
$\lvert\beta_2\text{-}1\rvert$	272.04	.30***	226.02	.04	276.90	.36***	199.29	-.03
Calculation								
PAE (log)	263.15	.43***	263.62	-.01	272.74	.39***	233.80	-.05
λ (log)	264.21	.43***	260.53	-.02	272.11	.39***	241.62	-.04
$\lvert\beta_1\text{-}1\rvert$	263.62	.43***	261.39	-.05	271.55	.39***	241.67	-.07
$\lvert\beta_2\text{-}1\rvert$	260.78	.44***	266.75	.01	271.35	.39***	213.22	-.05
Applied problems								
PAE (log)	275.29	.43***	244.86	-.16***	263.24	.41***	239.34	-.20***
λ (log)	275.71	.43***	238.00	-.17***	266.83	.42***	245.14	-.14***
$\lvert\beta_1\text{-}1\rvert$	275.66	.44***	238.67	-.17***	266.74	.43***	245.16	-.16***
$\lvert\beta_2\text{-}1\rvert$	275.28	.43***	247.44	-.17***	264.77	.43***	215.83	-.16***

Note. df_1 = degrees of freedom for g factor. b_1 = main effect of g factor. df_2 = degrees of freedom for NLE measure. b_2 = main effect of NLE measure.

***$p < .001$.

Twin Analyses

The twin sample allows for comparison of correlations of monozygotic (MZ) twins, who share 100% of their DNA, and dizygotic (DZ) twins, who share on average 50% of their DNA in order to get an estimation of the

amount of variance accounted for by genetics, shared environment (factors that make siblings more similar to one another) and nonshared environment (factors that make siblings less similar to one another) and error. OpenMx, a software package in R, was used to conduct the twin analyses in order to get estimates of heritability, shared environment, and nonshared environment/error for the accuracy values at ages 12 and 15 (Neale et al., 2016). The estimates are listed in Table 4. The accuracy measures at age 12 were accounted for by nonshared environment/error (.51-.60) and genetics (.40-.49), and shared environment was not a contributing factor for any of the measures. In contrast, by age 15, performance on the number line was accounted for almost entirely by nonshared environment/error (.72-.96).

Table 4

Correlations Between MZ Pairs and DZ Pairs and Estimates of Genetic, Shared Environment, and Non-Shared Environment and Error for NLE Accuracy Measures for Ages 12 and 15

	12-year-old					15-year-old				
NLE accuracy	r_{MZ}	r_{DZ}	a^2	c^2	e^2	r_{MZ}	r_{DZ}	a^2	c^2	e^2
PAE (log)	.46	.13	.40 [.12, .52]	0	.60 [.48, .71]	.27	.20	.06 [0, .38]	.19 [0, .34]	.75 [.61, .87]
λ (log)	.52	.09	.45 [.25, .56]	0	.55 [.44, .68]	.28	.28	.00 [0, .36]	.28 [0, .37]	.72 [.62, .83]
$\lvert\beta_1\text{-}1\rvert$	.51	.10	.45 [.30, .57]	0	.55 [.43, .55]	.27	.25	.05 [0, .41]	.22 [0, .36]	.73 [.58, .85]
$\lvert\beta_2\text{-}1\rvert$	.55	.10	.49 [.32, .60]	0	.51 [.40, .64]	.05	.02	.04 [0, .18]	.002 [0, .15]	.96 [.82, 1.0]

Note. Values in brackets represent 95% confidence intervals.

Discussion

The debate about how to appropriately characterize performance on the NLE task has left an open question about how the theoretical stances reflected in the measurements differentially translate to prediction of math achievement. Does a method that describes the average error such as PAE predict math achievement better than a method designed to capture the logarithmic-to-linear shift such as the mixed log-linear model or a method designed to capture strategy use such as the cyclical power model (one-cycle or two-cycle)? The results of the analyses of this study provide several conclusions: 1) PAE, mixed log-linear model, one-cycle cyclical power model, and two-cycle cyclical power model are highly correlated with one another 2) The accuracy measures for each provide more predictive value for the Applied Problems subtest than the other math achievement measures when g is included as a predictor 3) Differences in behavior genetic estimates are not noted among the accuracy measures.

First, the high correlations among the accuracy measures, especially in the 12-year-old sample, are notable given theoretical differences between the measures. Although the one-cycle cyclical power model and two-cycle cyclical power model account for performance based on strategy use, and the PAE and logarithmic-to-linear shift do not, the measures were still highly correlated. The highest correlation between the mixed log-linear model and one-cycle cyclical power model may be due to a similarity in the predicted shape of responding, in which both models capture overestimation of spaces on the lowest end of the number line. Such high correlations, especially between the mixed log-linear model and one-cycle cyclical power model, indicate that the different measures are mostly capturing the same variation.

Although all measures of the NLE task were significantly correlated both cross-sectionally and longitudinally with all three math achievement measures, the relationship between the NLE task and two measures of math achievement were no longer significant once general intelligence was included as a predictor. We have two possible explanations for why performance on the NLE task predicts performance on the Applied Problems subtest once g is controlled but does not predict performance on the Fluency or Calculation subtests.

First, it is possible that internal magnitude representation only assists in performance on tests that are developmentally challenging for the participant. For example, in children in kindergarten through second grade, NLE task performance was predictive of accuracy on simple addition and subtraction problems (Kim & Opfer, 2017). However, in a separate study of slightly older children (first through third grade subjects), NLE task performance was not predictive of performance on a timed arithmetic fact test, but NLE task performance was predictive of general curriculum math achievement (Sasanguie et al., 2013). In the case of kindergarten children, completion of addition and subtraction problems would be complex given the age, but by adolescence, perhaps "complexity" refers to applying mathematical concepts to story problems. Magnitude representation may assist in the development of complex mathematical abilities at different ages, but the level of complexity is determined by developmental stage.

Alternatively, the significant prediction of performance on the NLE task for Applied Problems may be due to shared characteristics of the tasks such as proportional reasoning requirements and strategy use. The Applied Problems subtest requires the participants to perform proportional tasks such as identifying what a third of a quantity would be in a story problem. In addition, story problems in the Applied Problems subtest require the participants to choose relevant information before performing operations; this is a more complex task that requires some strategy for higher performance (Woodcock, McGrew, & Mather, 2001).

The phenotypic analyses also gave insight into the stability of the task from age 12 to age 15. In the 3-year interval, the correlation between task performance was relatively stable (PAE: .37, λ: .44, β_1: .45, β_2: .28). In a previous study of 5-year olds, the correlation between performance (measured by PAE) on a NLE task across 30 weeks of measurement was .41 for 1-100 endpoints, .46 for 1-10 endpoints, and nonsignificant for 1-20 endpoints (Muldoon et al., 2013). Although the present study captures individuals during a different developmental period, and the time interval is larger, the measurement stability is similar between our study and the 1-10 as well as the 1-100 NLE task performance in Muldoon et al. (2013). It is possible that the correlations would have been higher in the adolescents if the test-retest interval was shorter, thus indicating more stability in performance as children age into adolescents, but that was not possible to directly assess in the present study.

As in the phenotypic analyses, the behavior genetic analyses also did not show any differentiation between the measures despite theoretical differences. Genetics were hypothesized to be influential in all measures given the amount of variation that is typically predicted in cognitive variables (Polderman et al., 2015). At age 12, genetics explain a large portion of the variation in the NLE performance, but by age 15, genetics are no longer predictive. Instead at age 15, individual differences are predicted almost entirely by the nonshared environment and measurement error, which means that the individual's environment is driving the variation rather than other more predictable parts of the environment such as school. By age 15, most subjects were able to complete the task quite well, so any individual variation that is explained may be due to individual motivation.

We also hypothesized that a significant proportion of the variation in performance on the NLE task would be due to shared environment because of the environmental influences demonstrated by previous studies

(Dehaene et al., 2008; Kucian et al., 2011; Opfer & Siegler, 2007; Ramani & Siegler, 2008; Siegler & Ramani, 2009; Thompson & Opfer, 2008, 2016). Contrary to our hypothesis, shared environment was not a significant factor. This indicates that environmental effects that make twins more similar to one another (e.g., shared curriculum) are not operating on individual differences performance on the NLE task at ages 12 and 15 despite evidence from other studies to suggest that the environment is an important component in performance. One explanation for the lack of shared environmental influences in the present sample is that shared environmental influences are reduced in cases where the environment is the same for participants (e.g. standardized curriculum) and thus cannot drive variation in the measure. In those situations, shared environmental factors such as schooling can affect the mean performance but may not lead to individual differences. Even so, behavior genetics results are in contrast to the hypothesized results for the logarithmic-to-linear shift, which would have predicted the shared environment to largely explain variation.

Overall, there do not appear to be fundamental differences between the accuracy measures on the NLE task in samples of 12 and 15 year olds. All measures are highly correlated and approximately equally predictive of math achievement. The appropriateness of the accuracy measure for a given study of adolescents thus can be determined based on pragmatic and theoretical underpinnings of the study. PAE is beneficial in that it is calculated without fitting data to a model, and thus even when subjects' responses are extreme, a result can still be obtained. However, if individual differences of progress towards linearity are being sought, then the mixed log-linear model seems to still be the most appropriate, although the fit may not be appropriate in the cases of very low performers. In addition, the similarity of the one-cycle cyclical power model with mixed log-linear model has also been established. The high correlation between these measures indicates that they are both measuring a similar pattern of responding, but the cyclical power model may lose individual differences for an even larger subset of the lowest performers due to model fit concerns. This study provided evidence for relevant individual differences in magnitude representation for adolescents, a group whose magnitude representation has not been largely studied. The similarity of the measures despite differences in theoretical underpinnings has also been shown using both regression and behavior genetic analyses.

Funding

The Western Reserve Reading and Math Project was supported by the Eunice Kennedy Shriver National Institute of Child Health and Development grants HD038075, HD059215, HD068728 and HD075460 and by National Center for Advancing Translational Sciences, grant 8UL1TR000090-05. S. Lukowski was supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1343012.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition, 122*, 454-459. doi:10.1016/j.cognition.2011.11.006

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

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

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48. doi:10.18637/jss.v067.i01

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

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

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 7867-7872. doi:10.1073/pnas.1402785111

Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic Bulletin & Review, 18*, 331-338. doi:10.3758/s13423-011-0059-z

Dackermann, T., Huber, S., Bahnmueller, J., Nuerk, H.-C., & Moeller, K. (2015). An integration of competing accounts on children's number line estimation. *Frontiers in Psychology, 6*, Article 884. doi:10.3389/fpsyg.2015.00884

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1-42. doi:10.1016/0010-0277(92)90049-N

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*, 371-396. doi:10.1037/0096-3445.122.3.371

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science, 320*, 1217-1220. doi:10.1126/science.1156540

Duncan, E. M., & McFarland, C. E. (1980). Isolating the effects of symbolic distance, and semantic congruity in comparative judgments: An additive-factors analysis. *Memory & Cognition, 8*, 612-622. doi:10.3758/BF03213781

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology, 123*, 53-72. doi:10.1016/j.jecp.2014.01.013

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47*, 1539-1552. doi:10.1037/a0025510

Goldsmith, H. H. (1991). A zygosity questionnaire for young twins: A research note. *Behavior Genetics, 21*, 257-269. doi:10.1007/BF01065819

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review, 107*, 500-524. doi:10.1037/0033-295X.107.3.500

Hurst, M., Leigh Monahan, K., Heller, E., & Cordes, S. (2014). 123s and ABCs: Developmental shifts in logarithmic-to-linear responding reflect fluency with sequence values. *Developmental Science, 17*, 892-904. doi:10.1111/desc.12165

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston naming test*. Philadelphia, PA, USA: Lea and Febiger.

Kenny, D. A., Kashy, D. A., Cook, W. L., & Simpson, J. A. (2006). *Dyadic data analysis*. New York, NY, USA: Guilford.

Kim, D., & Opfer, J. E. (2017). A unified framework for bounded and unbounded numerical estimation. *Developmental Psychology, 53*, 1088-1097. doi:10.1037/dev0000305

Kucian, K., Grond, U., Rotzer, S., Henzi, B., Schönmann, C., Plangger, F., . . . von Aster, M. (2011). Mental number line training in children with developmental dyscalculia. *NeuroImage, 57*, 782-795. doi:10.1016/j.neuroimage.2011.01.070

Lukowski, S. L., Soden, B., Hart, S. A., Thompson, L. A., Kovas, Y., & Petrill, S. A. (2014). Etiological distinction of working memory components in relation to mathematics. *Intelligence, 47*, 54-62. doi:10.1016/j.intell.2014.09.001

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*, 250-257. doi:10.1037/a0028240

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., . . . Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*, 535-549. doi:10.1007/s11336-014-9435-8

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*, 169-195. doi:10.1016/j.cogpsych.2006.09.002

Opfer, J. E., Thompson, C. A., & Kim, D. (2016). Free versus anchored numerical estimation: A unified approach. *Cognition, 149*, 11-17. doi:10.1016/j.cognition.2015.11.015

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*, 117-134. doi:10.1007/s10212-015-0251-z

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderheiser, J. (2013). *Behavioral genetics*. New York, NY, USA: Worth Publishers, Macmillan Learning.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics, 47*, 702-709. doi:10.1038/ng.3285

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*, 375-394. doi:10.1111/j.1467-8624.2007.01131.x

Rouder, J. N., & Geary, D. C. (2014). Children's cognitive representation of the mathematical number line. *Developmental Science, 17*, 525-536. doi:10.1111/desc.12166

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology, 30*, 344-357. doi:10.1111/j.2044-835X.2011.02048.x

Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology, 114*, 418-431. doi:10.1016/j.jecp.2012.10.012

Sasanguie, D., & Reynvoet, B. (2013). Number comparison and number line estimation rely on different mechanisms. *Psychologica Belgica, 53*, 17-35. doi:10.5334/pb-53-4-17

Sasanguie, D., Verschaffel, L., Reynvoet, B., & Luwel, K. (2016). The development of symbolic and non-symbolic number line estimations: Three developmental accounts contrasted within cross-sectional and longitudinal data. *Psychologica Belgica, 56*, 382-405. doi:10.5334/pb.276

Semel, E., Wiig, E. H., & Second, W. A. (2003). *Clinical evaluation of language fundamentals: Examiner's manual* (4th ed.). San Antonio, TX, USA: Harcourt Assessment.

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

Siegler, R. S., & Lortie-Forgues, H. (2014). An integrative theory of numerical development. *Child Development Perspectives, 8*, 144-150. doi:10.1111/cdep.12077

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-243. doi:10.1111/1467-9280.02438

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*, 545-560. doi:10.1037/a0014239

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain and Education: The Official Journal of the International Mind, Brain, and Education Society, 3*, 143-150. doi:10.1111/j.1751-228X.2009.01064.x

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General, 142*, 193-208. doi:10.1037/a0028560

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 683-692. doi:10.1037/0096-1523.16.4.683

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology, 101*, 20-51. doi:10.1016/j.jecp.2008.02.003

Thompson, C. A., & Opfer, J. E. (2016). Learning linear spatial-numeric associations improves accuracy of memory for numbers. *Frontiers in Psychology, 7*, Article 24. doi:10.3389/fpsyg.2016.00024

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *The comprehensive test of phonological processing — Examiner's manual*. Austin, TX, USA: PRO-ED.

Wechsler, D. (2004). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX, USA: Harcourt Assessment.

Wood, G. M., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science Quarterly, 50*, 489-525.

Woodcock, R. W. (1998). *Woodcock reading mastery tests — Revised examiner's manual*. Circle Pines, MN, USA: American Guidance Service.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL, USA: Riverside Publishing.