Journal of Numerical Cognition

PsychOpen GOLD

Empirical Research

Optimizing the 0-100 Number Line Estimation Task: Scale Reduction and Its Implications for Elementary Mathematical Cognition

Kamal Chawla[1], Julie L. Booth[2], Christina Areizaga Barbieri[3]

[1] *College of Education and Human Development, University of Maine, Orono, ME, USA.* **[2]** *College of Liberal Arts, Temple University, Philadelphia, PA, USA.* **[3]** *College of Education and Human Development, University of Delaware, Newark, DE, USA.*

Journal of Numerical Cognition, 2026, Vol. 12, Article e17459, https://doi.org/10.5964/jnc.17459

Received: 2025-03-26 • **Accepted:** 2025-12-31 • **Published (VoR):** 2026-05-08

Handling Editor: Darren J. Yeo, Nanyang Technological University, Singapore, Singapore

Corresponding Author: Kamal Chawla, 326 Shibles Hall, College of Education and Human Development, University of Maine, Orono, United States. E-mail: kamal.chawla@maine.edu

Abstract

We investigate the optimal number of items for the 0-100 number line estimation task used in research on children's mathematical cognition and learning. In this paper, we reanalyzed data involving $N = 234$ students, applying an Item Response Theory- Graded Response Model to identify items with high discrimination parameters (> 1.0), iteratively reducing the 23-item scale by including items with discrimination values close to 1.0 until the reduced scale produced comparable scores to the original. Our analysis identified a reduced scale of 15 items that maintained strong correlations with–and produced consistent patterns of developmental change and predictive capability compared to–the original scale. Our findings demonstrate that a reduced 0-100 number line estimation task can effectively measure numerical magnitude understanding (accuracy and linearity of estimates) from kindergarten through third grade while saving time and resources.

Keywords

mathematical cognition, mathematics, number line, estimation, cognitive development, reduced scale, item response theory

Over the past two decades, the number line estimation (NLE) task has arguably become one of the most well-known methods of measuring numerical competence in adults and children alike. A simple Google Scholar search using the term "Number Line Estimation" yielded over 3300 hits, and the seminal NLE papers have each been cited over 1100 times (Booth & Siegler, 2006, *1102 citations*; Siegler & Booth, 2004, *1566 citations*). Though each of these citations certainly does not indicate that the task itself was used in the study, it is clear, at a minimum, that results from NLE tasks have been widely used to shape our understanding of numerical cognition.

The core NLE task involves presenting a series of blank number lines with the endpoints of the line marked by specific numbers; for each item, a third number that is numerically somewhere between the two endpoint numbers is given, and the participant's task is to estimate where that number would fall on the number line (e.g., Siegler & Booth, 2004). For example, on a 0-100 number line task, the number "0" would be marked at the left endpoint and "100" marked at the right endpoint, and an individual item might ask participants to demonstrate where they believe individual numbers (e.g., 23, 47, 2, 96, 14, etc.) belong on the line. Scoring of this task typically involves either computing a participant's estimation accuracy across all of their items

$$\text{Percent of Absolute Error (PAE)} = \left| \frac{\textit{Target Number} - \textit{Estimated Number}}{\textit{Scale of Estimates}} \right| \times 100$$

(Siegler & Booth, 2004) or examining a participants' pattern of estimates across all of their items by fitting different types of models to their estimates (e.g., linear, logarithmic, exponential (Booth & Siegler, 2006); power models (Barth & Paladino, 2011; Ruiz et al., 2023); mixed log linear models (Qin et al., 2024)).

The use of the NLE task has been widespread across both contexts and cultures. For example, the task has been administered in laboratory or one-on-one settings (Praet & Desoete, 2014; Siegler & Booth, 2004; Wall et al., 2016), online using platforms such as MTurk (Landy et al., 2017) and Qualtrics (Fitzsimmons & Thompson, 2022), in an FMRI scanner (Vogel et al., 2013), and in both K-12 (Barbieri et al., 2021; Jung et al., 2020) and college classrooms (Steinke, 2017). NLE tasks have also been employed in studies across the globe, including in Israel (Ashkenazi & Cohen, 2023), Singapore (Ruiz et al., 2024), Chile (Xu et al., 2023), Luxembourg (Nuraydin et al., 2023), Zambia (Sudo et al., 2022), China (Li et al., 2024), Germany (Jung et al., 2020), Turkey (Sarı & Olkun, 2021), and with individuals from Indigenous tribes (Dehaene et al., 2008).

The characteristics of the number line task have varied widely across studies. For example, many studies use a paper-and-pencil administration (e.g., Booth & Siegler, 2006; Chan & Mazzocco, 2024), while others use computerized versions of the task (e.g., Booth & Siegler, 2008; Fazio et al., 2014; Gunderson & Hildebrand, 2021). Most studies use the number-to-position version of the task described above, though others have given a mark on a number line and asked the participant to estimate what number goes there

(Position-to-Number, e.g., Peeters et al., 2017; Siegler & Booth, 2004). And while most studies present each item individually as described above (i.e., participants only see one number line at a time), others have provided a page containing several number lines that are visible at once (Barbieri et al., 2023; Young & Booth, 2015) or have asked participants to place multiple numbers on the same number line (Siegler & Booth, 2004; Steinke, 2017).

Perhaps the greatest variability across studies using number line estimation tasks is the numerical scale and the type of numbers to be placed. While many studies have continued to employ the 0 – 100 or 0 – 1000 whole-number scales introduced in the original studies (e.g., Hoard et al., 2008; Praet & Desoete, 2014; Sarı & Olkun, 2021; Sullivan et al., 2011), others have used much smaller (e.g., 0 – 10, Dietrich et al., 2016; 0 – 20, Cornu et al., 2017; 0 – 8, Yu et al., 2022) or larger scales (e.g., 0 – 10,000, Jung et al., 2020; 0 – 1,000,000, Slusser et al., 2013), irregular scales such as 0 – 62,571 (Booth et al., 2014) or 1,000 – 1,000,000,000 (Landy et al., 2017), and scales involving negative numbers (-1000 – 0, Brez et al., 2016; -1,000 – 10,000, Young & Booth, 2015); some have even explored the use of unbounded number lines (e.g., Link et al., 2014; Reinert et al., 2019). The number line task has also been used to measure participants' understanding of different types of numbers, including fractions (e.g., Booth & Newton, 2012; Namkung & Fuchs, 2016), decimals (DeWolf et al., 2015; Schneider et al., 2009), and percentages (Schiller et al., 2024).

Despite the wide variability in the particulars of NLE task administration, there is a large consensus that individuals' NLE performance is related to (and often predictive of) their general mathematical competence. For example, NLE has been shown to predict arithmetic performance (e.g., Dietrich et al., 2016; Gunderson & Hildebrand, 2021), problem-solving skills (e.g., Zhu et al., 2017), and mathematical reasoning (e.g., Ruiz et al., 2024). In a meta-analysis of 41 papers, Schneider and colleagues (2018) reported strong associations between NLE and counting, computation skills, and school mathematics achievement. Ellis and colleagues (2021) replicated these overall findings, emphasizing the importance of performance on the NLE task, but found performance to be particularly predictive of mathematical competence in younger children. NLE performance even correlates with brain activation while solving arithmetic problems (Berteletti et al., 2015), and poor NLE performance has been linked to mathematical learning disabilities (Geary et al., 2012). In a recent systematic review of 33 manuscripts that examine predictors of algebra performance, NLE tasks were one of the most studied student-level factors, with fraction NLE tasks consistently demonstrating their predictive utility, but whole number line tasks were also often used (Silla et al., under review).

The prevalence of its employment and the usefulness of the task for predicting more complex mathematical competencies make it likely that the NLE will continue to be an essential measure in psychology and education research. Given its prominence as a predictor of school mathematics achievement and the fact that NLE data collection often

takes place in schools during precious classroom time—either with full-class administration during class or by pulling individual students out of class—it is necessary to ensure that the task is as efficient as possible. However, the number of target items has varied quite a bit from study to study, ranging from a minimum of 6 items to 44 or more items (Schneider et al., 2018). Additionally, although studies with older children and adolescents often use paper-based tasks that can be administered reasonably quickly in whole-group settings (e.g., Barbieri et al., 2021), this is not the case for younger children. In studies with young children, number line tasks are often administered one-on-one with an experimenter using an iPad (e.g., Geary et al., 2008). Given that the 0 – 100 scale, in particular, is among the most widely used scales (696 Google Scholar hits, compared with 465 for 0 – 1000 and 542 for 0 – 1) and is useful for young children (who likely do not yet have fully developed attentional control; Tremolada et al., 2019), we aim to determine how few target items could be used to get an accurate measure of children's numerical magnitude understanding to ensure responsible use of teachers' and students' time.

Scale Reduction Techniques

Scale reduction techniques are widely used in research to create more efficient versions of longer measurement scales while maintaining their essential psychometric properties. These methods are especially valuable when researchers seek to reduce participant burden, streamline data collection, and increase the feasibility of applying the scale in various settings.

One technique used in this study is the discrimination parameter method, derived from the Graded Response Model (GRM) within Item Response Theory (IRT). The GRM is particularly suited for scales with ordered categories, such as the number line estimation tasks in the present study. By prioritizing items with high discrimination parameters, researchers can retain items that are most effective in differentiating between individuals with similar underlying traits (Embretson & Reise, 2000). Discrimination parameters reflect how well an item distinguishes between respondents at different levels of the measured latent trait. Items with low discrimination values are discarded since they contribute little to the overall precision of the scale.

The optimal range of discrimination parameters in GRM typically falls between 0.5 and 2 (Hambleton et al., 1991). Items with discrimination values in this range are considered highly effective at differentiating between respondents, making them ideal candidates for retention on a reduced scale. Lower values, particularly below 0.5, indicate weak discrimination, and such items are often removed from the scale due to their minimal contribution to measurement accuracy. Items with high discrimination values (above 1.5) are particularly effective at distinguishing between respondents with different ability levels, while items with moderate discrimination values (between 0.5 and 1.5) still contribute meaningfully to the overall precision of the scale. Setting a value around

1.0 ensures that only items with optimal discrimination remain on the reduced scale. This technique allows for a more concise instrument without compromising its ability to measure the intended construct.

Another common method for scale reduction focuses on assessing model fit indices, such as Root Mean Square Error (RMSE) and Standardized Root Mean Square Residual (SRMR). These indices are frequently used in confirmatory factor analysis (CFA) to evaluate how well a model (with a reduced number of items) fits the data. While RMSE and SRMR help provide an overall picture of model fit, they are often used in conjunction with discrimination-based approaches to ensure both psychometric robustness and conceptual clarity of the reduced scale (Brown, 2015). In these approaches, researchers typically retain items that contribute to lower error rates and higher overall fit indices, making them essential tools when developing shorter scales that maintain their integrity.

Lastly, some techniques focus on maximizing the amount of information provided by the scale at different points on the trait continuum. In these approaches, often called test information functions, items are selected based on their contribution to the total information at key points (e.g., low, moderate, and high levels of the trait). This method allows researchers to ensure that the reduced scale provides reliable measurements across the entire range of trait levels (Baker, 2001).

By employing discrimination parameters through the GRM and prioritizing items with high thresholds, the current study leverages an efficient method to produce a reduced scale that remains effective and valid for measuring developmental changes in number line estimation tasks. This technique is balanced against other reduction strategies to ensure that the final scale continues to perform comparably to the original instrument.

The Present Study

In the present study, we investigate the usefulness of scale reduction techniques for finding the optimal number of items for the 0 – 100 number line estimation task. By reanalyzing the original data from two seminal studies that employed this measure (Booth & Siegler, 2006; Siegler & Booth, 2004), we aim to determine the minimal set of items that could produce a scale with the same psychometric properties as the original, longer scale. These seminal studies employed the 0 – 100 number line task with students from kindergarten to second grade and kindergarten to third grade, respectively. We recognize that second and third graders are typically quite linear and accurate on the 0 – 100 number line scale, and thus number line estimates on the 0 – 1000 scale might be a more appropriate indicator of their numerical competence. However, we feel that investigating the potential of a reduced scale on this most widely used scale (0 – 100) requires demonstration of replication of the findings from the original papers, which would require the use of all the data included in the original papers, not just that of kindergartners and first graders for whom this scale might be most relevant. We thus

compare potentially reduced scales to the original scale on two measures of interest from the original studies to determine if a reduced scale yields sufficiently comparable performance scores for students. We conduct the exploratory factor analysis (EFA) to draw meaningful psychometric comparisons between the two scales. Finally, we aim to replicate key findings from the original studies to determine if a reduced scale would yield the same results as the original studies' research questions around developmental changes in number line estimates and the correlation of those estimates with mathematics achievement.

Method

Data Sources

The dataset comprises accumulated data from three previous studies: Siegler and Booth (2004) Experiment 1; Siegler and Booth (2004) Experiment 2; and Booth and Siegler (2006) Experiment 1, each of which measured number line estimation on a 0-100 scale using 23 common number items: 3, 4, 6, 8, 12, 17, 21, 24, 25, 29, 33, 39, 48, 52, 57, 61, 64, 72, 79, 81, 84, 90, and 96; the Siegler and Booth (2004) study also used 43, while the Booth and Siegler (2006) study used 42; these items were therefore eliminated from the analysis. Estimates for these 23 numbers, grade level, age, school, and achievement score, were compiled from $N = 234$ students across the three studies: 61 kindergartners, 76 first-graders, 75 second-graders, and 22 third-graders. In this task, students were presented with sheets of paper, one at a time, each containing a 25 cm number line with a 0 marked at the left endpoint and 100 marked at the right endpoint. A number between 0 and 100 was printed at the top center of the page, and the student's task was to indicate where on the number line they believed that number would go; they did this by making a mark on the number line at the preferred spot. For each item, the original researchers measured the number of millimeters from the left endpoint to the student's placement. This was divided by the length of the whole scale and multiplied by 100 to determine what number should go where the student placed the mark. The data used for the present study consisted of those numbers for each student for each item and the student's age, grade level, and mathematics achievement score.

Assessment Framework

This study utilizes number line estimation as a diagnostic tool for evaluating students' numerical magnitude representation capabilities. This approach is integral to discerning foundational mathematical understanding, necessitating an exploration of student response patterns to integer estimations on a numerical line. The investigation encompasses the evaluation criteria, the re-categorization of responses, and the probability

assessment of endorsing a response predicated on item difficulty and discrimination attributes.

Item Selection Criteria

The redevelopment of the number line estimation scale was guided by the utilization of high discrimination parameters and an optimum threshold value (~1.0), as recommended by Embretson and Reise (2000) and Hambleton and colleagues (1991). This method prioritizes the differentiation capacity of items based on their ability to distinguish between individuals of marginally differing trait levels, coupled with the spacing and meaningfulness of response categories. Having said that, using this method, we can refine the scale by discarding the previously developed items whose discrimination parameter is significantly lower than the threshold value of 0.1. This approach underpins the construction of a concise yet efficacious assessment tool, contrasting with alternative methodologies that emphasize model fit and error metrics, such as RMSE and SRMR values.

Grading System and Data Transformation

The grading system operationalized in this study delineates responses within ±5 points from the correct answer as '0' (very close), beyond 5 points as '1' (far), and beyond 10 points as '2' (very far). Inversely, responses below -5 points were categorized as '-1' (far), and those below -10 points as '-2' (very far). The ±5/±10 cutoffs were selected for interpretability and because they yielded ordered thresholds and well-behaved category response curves without sparse categories; alternative nearby cutoffs (±3/±6, ±5/±15, ±3/±7) produced substantively similar item discriminations and test information, supporting robustness.

We computed signed estimation error on the 0 − 100 scale and transformed responses to magnitude-only ordered categories required by a unidimensional graded response model: 3 = very close ($|\text{error}| \leq 5$), 2 = far ($5 < |\text{error}| \leq 10$), 1 = very far ($|\text{error}| > 10$). Here, it is important to note that points were transformed to a new scale of 1 − 3 to align with the GRM's unidimensional, ordered-category assumptions; the numeric labels are ordinal, and the GRM operates on the latent trait via estimated discrimination and threshold parameters.

Probability of Endorsing a Response

Two pivotal factors inform a student's likelihood of selecting a specific response category, reflecting their deviation from the accurate estimation: item difficulty and discrimination. The polytomous nature of item responses, ranging from very close to very far from the correct answer, necessitates the application of the IRT-Graded Response Model (IRT-GRM) for parameter estimation. This model facilitates the nuanced analysis of the

endorsement probabilities, yielding insights into each question's underlying challenge and discriminatory power. The GRM requires an ordered set of response categories that reflect increasing deviations from the correct placement; the cutoffs here define an ordinal, severity-graded outcome (very close ⟶ very far) on each item, which the GRM models via item slopes (discrimination) and ordered threshold parameters that separate adjacent categories.

Data Preparation

To conduct the rest of the planned analyses, we first computed two types of performance scores for each individual student based on each iteration of the reduced scale.

Accuracy of Estimates

We used Percent Absolute Error (PAE; e.g., Siegler & Booth, 2004) to measure the overall accuracy of students' estimates on the number line

$$\text{PAE} = \left| \frac{\textit{Target Number} - \textit{Estimated Number}}{\textit{Scale of Estimates}} \right| \times 100.$$

PAE is thus computed for a given child on a given item by subtracting the actual value of the to-be-estimated number from the numerical value that corresponds with the child's placement on the number line for that number, taking the absolute value so that over- vs. underestimates are not distinguished, and dividing the value by the scale of the number line (in this case, 100 since it is a 0 – 100 number line); the resulting value is then multiplied by 100 to get a percentage. The average PAE is then computed across all the items for a given child to provide a measure of overall accuracy. For the present study, PAE scores were computed for each child for the entire scale and then, in turn, for each potential reduced scale.

Pattern of Estimates

As in Siegler and Booth (2004) and Booth and Siegler (2006), we first computed the median estimate for each number of children in a particular grade and fit linear and logarithmic functions to the median estimates at each grade level. We then recorded the variance that could be explained by the best-fitting linear (R^2_{lin}) and logarithmic function (R^2_{log}) at that grade level. This process was undertaken across the entire scale and for each potential reduced scale.

Then, to obtain a measure of the pattern of individual students' estimates, we fit linear and logarithmic functions to each child's estimates and recorded the amount of variance that could be explained by the best-fitting linear (R^2_{lin}) and logarithmic function (R^2_{log}) for that child's estimates. Again, these values were computed for each child for the entire scale and then for each potentially reduced scale.

Exploratory Factor Analysis (EFA)

To further validate the reduced scale, we conducted Exploratory Factor Analysis (EFA) on both the original and reduced scales using the *psych* package in R (Revelle, 2021). We performed parallel analysis to determine the number of factors to retain (Horn, 1965). To assess unidimensionality, we calculated the eigenvalue ratio (first to second factor) and the percentage of variance explained by the first factor (Slocum-Gori & Zumbo, 2011). We also assessed reliability using Cronbach's alpha and McDonald's omega (Dunn et al., 2014) for further comparisons.

Results

Graded Response Model (GRM) Analysis

After running the Item Response Theory (IRT)-Graded Response Model using the *ltm* package in R (Rizopoulos, 2018) for our polytomous dataset, we found threshold (extreme) and discrimination values, as mentioned in Table 1. In the Graded Response Model, thresholds indicate where individuals are likely to transition between different response categories, with higher thresholds corresponding to more difficult items or response categories that require a higher level of the latent trait.

As mentioned previously, discrimination parameters indicate that an item is effective at differentiating between individuals with slightly different trait levels. Therefore, for our first iteration, we considered all the items whose discriminating values are more significant than 1 and then subsequently added items that are very close to 1, one by one, for further iterations. We stopped at the reduction iteration that yielded a comparative analysis that produced similar results to those of the original dataset.

As indicated in Table 1, all the values greater than target item 33 have low discrimination values less than 1. Thus, we reduced the length of the test by removing all the target items after 33 for this scale for the first iteration (the remaining items are 3, 4, 6, 8, 12, 17, 21, 24, 25, 29, 33). We then successively added back in one item at a time, starting with the largest discrimination value (i.e., we subsequently added 52 for the second iteration (discrimination value 0.997), 39 and 57 for the third (discrimination value 0.88), etc.), and 81 for the fourth (discrimination value 0.83). Thus, the original combined scale had 23 items, Iteration 1 had 11 items, Iteration 2 had 12 items, Iteration 3 had 14 items, and Iteration 4 had 15 items. We did not proceed with Iteration 5 because no further items had discrimination values close to the threshold of 1.0.

Table 1

Threshold and Discrimination Values From Graded Response Model Analysis

Target Number	Threshold 1	Threshold 2	Discrimination
3	-0.787	0.125	2.182
4	-0.048	0.771	2.189
6	0.284	1.002	2.421
8	0.684	1.174	2.907
12	0.610	1.292	2.499
17	0.735	1.475	2.155
21	0.650	1.421	1.788
24	0.579	1.454	1.318
25	0.405	1.369	1.837
29	0.605	1.446	1.173
33	0.302	1.081	1.246
39	0.433	1.635	0.881
48	0.987	6.014	0.341
52	-0.393	0.634	0.997
57	-0.131	0.982	0.880
61	-0.630	0.969	0.554
64	-0.399	1.207	0.598
72	-0.791	1.332	0.517
79	-0.047	2.557	0.472
81	0.071	1.687	0.837
84	0.702	3.201	0.384
90	0.111	1.747	0.521
96	1.999	5.182	0.341

Student Performance

To test whether the reduced scale options yield sufficiently comparable performance scores for students, we first computed correlation coefficients for the PAE scores for the original scale with the PAE scores for each potential reduced scale and for the R^2_{lin} values for the original scale with those for each potential reduced scale.

PAE scores for the original scale were significantly correlated with PAE scores for each of the potentially reduced scales as shown in Table 2. To compare these correlations, we used Fisher's *r*-to-*z* tests for dependent samples. Accuracy for Iteration 2 was more highly correlated with that for the original scale than was Iteration 1 accuracy. Iteration 3 accuracy was more highly correlated than Iteration 2 accuracy, and Iteration 4 accuracy was more highly correlated with Iteration 3 accuracy. Thus, of all the potentially reduced scales, Iteration 4 produced the most highly correlated accuracy scores with those from the original scales. Similarly, R^2_{lin} scores for the original scale were significantly correla-

ted with R^2_{lin} scores for each of the potentially reduced scales, as presented in Table 3, and we found similar results.

Table 2

PAE Scores: Original vs Iterations

Comparison	Correlation Coefficient (*r*)	Degrees of Freedom (*df*)	*p*	Fisher's *z*	*z*	*p* (*z*-test)
Original vs. Iteration 1	0.746	234	< .001	–	–	–
Original vs. Iteration 2	0.756	234	< .001	–	–	–
Original vs. Iteration 3	0.776	234	< .001	–	–	–
Original vs. Iteration 4	0.797	234	< .001	–	–	–
Iteration 1 vs. Iteration 2	–	–	–	-3.642	–	< .001
Iteration 2 vs. Iteration 3	–	–	–	-4.741	–	< .001
Iteration 3 vs. Iteration 4	–	–	–	-5.182	–	< .001

Table 3

R^2_{lin} *Scores Comparisons*

Comparison	Correlation Coefficient (*r*)	Degrees of Freedom (*df*)	*p*	Fisher's *z*	*z*	*p* (*z*-test)
Original vs. Iteration 1	0.462	234	< .001	–	–	–
Original vs. Iteration 2	0.736	234	< .001	–	–	–
Original vs. Iteration 3	0.814	234	< .001	–	–	–
Original vs. Iteration 4	0.926	234	< .001	–	–	–
Iteration 1 vs. Iteration 2	–	–	–	-8.771	–	< .001
Iteration 2 vs. Iteration 3	–	–	–	-5.829	–	< .001
Iteration 3 vs. Iteration 4	–	–	–	-13.858	–	< .001

Next, we computed a series of paired-sample *t*-tests to compare the mean PAE and R^2_{lin} scores for each potential reduced scale with the mean scores for the original scale, as shown in Table 4.

In Table 4, we summarize the *t*-test results for both PAE and R^2_{lin} measures across all four iterations of the reduced scale, including the *t*-values, degrees of freedom, *p*-values, Cohen's *d* effect sizes, and interpretation of effect sizes. To account for multiple comparisons, we applied the Bonferroni correction, adjusting the significance threshold to $\alpha = .05/8 = .00625$. Only one comparison yielded a statistically significant difference after this correction (that for Iteration 1 PAE). PAE and R^2_{lin} for Iterations 2, 3, and 4 did not significantly differ from the original scale.

Table 4

Differences Between Each Iteration and Original Scale in PAE and R^2_{lin} Metrics

Measure	Iteration	*t*	*df*	*p*	*d*	Effect Size
PAE	1	-3.003	466	.003*	0.28	Small
	2	-2.451	466	.015	0.23	Small
	3	-1.906	466	.057	0.18	Small
	4	-1.727	466	.085	0.16	Small
R^2_{lin}	1	2.287	466	.023	0.21	Small
	2	2.296	466	.022	0.21	Small
	3	2.068	466	.039	0.19	Small
	4	1.956	466	.051	0.18	Small

Note. To account for multiple comparisons, we applied a classic Bonferroni correction, adjusting the significance threshold to $\alpha = .05/8 = .00625$.
*$p < .00625$.

Selecting a Preferable Reduced Scale

We considered the findings from the prior analysis and used them to determine which iteration we deem most preferable. Though Iteration 1 was the only reduced scale that had discrimination values all of less than one, this scale also had all eleven target items that were focused on the lower range of the 0 – 100 scale, specifically between 3 and 33 (see Table 1). Thus, we were concerned that this scale would not be able to fully capture any developmental changes that occurred at the higher end of the scale. We decided that a reduced scale that also represented the higher end of the 0 – 100 range would be most preferable. Thus, we considered Iterations 2 – 4. Iterations 2 and 3 added target items at around the midpoint of the scale (i.e., 52 for Iteration 2 and 39 and 57 for Iteration 3). Thus, we saw these as improvements upon Iteration 1. However, Iteration 4 included a target item at the higher end of the scale (i.e., 81) and we saw this as a particular benefit to the scale. We also considered correlations between the original scale and reduced scales' R^2_{lin} and PAE scores for each iteration. As displayed in Table 3, we found significant correlations for all iterations, yet correlations for our R^2_{lin} comparisons became significantly and dramatically stronger as we progressed from Iteration 1 (r = .462) to Iteration 4 (r = .926). As displayed in Table 2, correlations between original and reduced scale PAE scores also became progressively and significantly stronger, though the increases were not as drastic, going from Iteration 1 (r = .746) to Iteration 4 (r = .797).

Next, we considered differences in original PAEs and R^2_{lin} to each of the reduced scales. As displayed in Table 4, considering the Bonferroni-corrected p-value threshold of .00625, only one significant difference was found: PAE for Iteration 1 was significantly

different from PAE for the original scale. R^2_{lin} for Iteration 1, as well as PAE and R^2_{lin} for Iterations 2, 3, and 4, were not significantly different from that on the original scale.

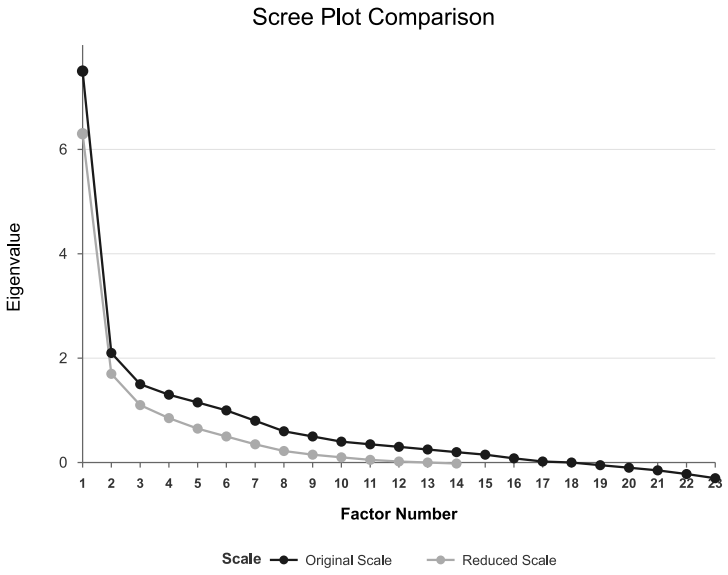
Based on these analyses, we recommend Iteration 4 as the preferred reduced scale, as it produced the PAE and R^2_{lin} values that are most highly correlated with the original scale, and because those PAE and R^2_{lin} values were not significantly different from the original scale. We thus used Iteration 4 in the subsequent analyses to determine if key findings from the original studies are replicated with the reduced scale. This scale included the following 15 items: 3, 4, 6, 8, 12, 17, 21, 24, 25, 29, 33, 39, 52, 57, and 81.

Exploratory Factor Analysis (EFA) Analysis

Before replicating key findings from the original studies, we conducted Exploratory Factor Analysis (EFA) to examine the underlying factor structure of both the original and reduced scales and to assess their dimensionality. We first examined the factor structure using principal axis factoring with oblimin rotation. Scree plots were generated to visually compare the eigenvalues of both scales, as shown in Figure 1.

Figure 1

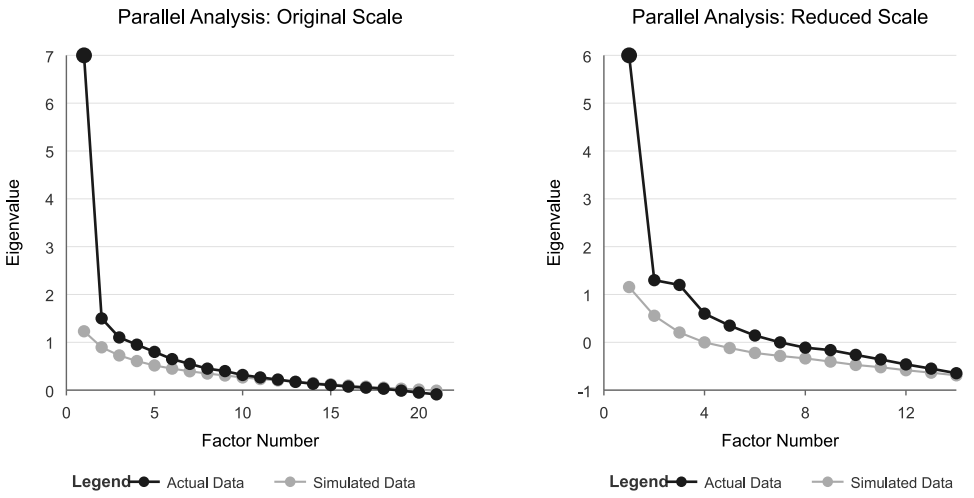
Scree Plot Comparison for Both Original Scale and Reduced Scale



Parallel analysis was then performed to determine the number of factors to retain (Horn, 1965), with results displayed in Figure 2.

Figure 2

Parallel Analysis Comparisons



We compared the observed eigenvalues with those generated from random data to identify significant factors. To assess unidimensionality, we calculated the eigenvalue ratio (first to second factor) and the percentage of variance explained by the first factor (Slocum-Gori & Zumbo, 2011). Reliability was assessed using Cronbach's alpha and McDonald's omega (Dunn et al., 2014). Additionally, we examined factor loadings to compare the clarity of factor structure between the original and reduced scales.

As shown in Table 5, the reduced scale maintained comparable psychometric properties to the original scale while improving some aspects of unidimensionality. The reduced scale (15 items) demonstrated a slightly lower eigenvalue ratio (3.64) compared to the original scale (3.80), but explained a higher percentage of variance through its first factor (56.49% vs. 54.85%). Both scales exhibited high internal consistency reliability, with the original scale showing marginally higher Cronbach's alpha (α = .86 vs. .85) and identical McDonald's omega values (ω = .89). Notably, parallel analysis suggested fewer factors for the reduced scale (2 factors) compared to the original scale (4 factors), indicating a potentially simpler factor structure. These results suggest that the reduced scale preserves the essential psychometric qualities of the original while potentially offering a more parsimonious measure of the construct.

Table 5

Psychometric Comparisons

Property	Original Scale	Reduced Scale
Number of Items	23	15
Eigenvalue Ratio (Factor 1/Factor 2)	3.82	3.64
Variance Explained by First Factor (%)	54.85	56.49
Cronbach's Alpha	0.86	0.85
McDonald's Omega	0.89	0.89
Factors Suggested by Parallel Analysis	4	2

Note. Reduced Scale: 3, 4, 6, 8, 12, 17, 21, 24, 25, 29, 33, 39, 52, 57, 81.

Developmental Changes in Accuracy of Number Line Estimates

As in the original studies, we first conducted a one-way ANOVA on PAE scores by grade. In the original publications, there were significant main effects of grade on PAE scores such that kindergartners' estimates were less accurate than those of first or second graders (Siegler & Booth, 2004) or first, second, and third graders (Booth & Siegler, 2006). With the reduced scale, there was also a significant main effect of grade, $F(3,230) = 52.766$, $p < .001$, $\eta_p^2 = .408$. Follow-up paired-sample t-tests with Bonferroni correction showed that kindergarteners had higher PAE ($M = .28$) than first ($M = .16$), second ($M = .13$), or third graders ($M = .09$); first graders also had significantly higher PAE than third graders (all $p < .05$).

Developmental Changes in Patterns of Number Line Estimates

As in the original studies, we used a three-pronged approach to examining developmental change in patterns of number line estimates: 1) Comparing variance explained by linear vs. logarithmic functions fitting median estimates at each grade level, 2) Comparing the distribution of the best-fitting type of function (Linear or Logarithmic) for individual students in each grade level, and 3) examining changes in the amount of variance accounted for by the best fitting linear function (R^2_{lin} scores) for each child by grade level.

As previously mentioned, since the original studies, there have been a number of efforts to test additional functions beyond the linear and logarithmic to assess the nature of the pattern of students' number line estimates. While it is not practical to test the effectiveness of the reduced scale using all of the types of models that have been introduced, there is one particular model, the mixed log linear model, which relies on a similar theoretical framework as the original studies and which has been shown to parsimoniously capture individual differences in most students' number line estimates (Qin et al., 2024). As one final test of the effectiveness of the reduced scale, we thus conclude this section by assessing change in the fit of the mixed log linear model to students' estimates on the original and the reduced scales.

Variance Explained by Linear vs. Logarithmic Functions Fitting Median Estimates

In the original studies, median estimates were computed for each number for students in a particular grade level. That is, the median estimate for the placement of 3 by kindergartners, the median estimate for the placement of 3 by first graders, the median estimate for the placement of 4 by kindergartners, etc. Then, for each grade, the median estimates were plotted against the actual values of the numbers, and the best-fitting linear and logarithmic functions were identified. The absolute value of the differences between the median estimates and the number predicted by the best fitting linear and logarithmic functions were then computed and compared. Results from those studies generally indicated that kindergartners' median estimates were better fit by the logarithmic function, first graders' estimations were typically equally well fit by the two functions, and second and third graders' estimations were better fit by the linear function (Booth & Siegler, 2006; Siegler & Booth, 2004)[1]. Plots showing the best-fitting linear and logarithmic functions at each grade level for the original and reduced scales can be found in Figure 3.

To evaluate whether a linear or logarithmic model was a better fit for the median values at each grade level, we first computed, for each trial at each grade level, the absolute value of the difference between the median value and the value for that trial that would be predicted by the best-fitting linear and logarithmic models. We then computed paired-samples *t*-tests for each grade level to compare the average distance of the median values to the linear vs. the logarithmic model; each of the included trials was therefore a separate data point in this analysis. Results for the original scale containing 23 trials, consolidated across studies, were aligned with those described in the original studies (see Figures 3a-d): Kindergartners' estimates were better fit by the logarithmic (R^2 = .93) than the linear function (R^2 = .67; $t(22)$ = 3.97, p < .001, d = 0.80); First graders were equally fit by the logarithmic (R^2 = .96) and linear functions (R^2 = .95) as indicated by a non-significant; $t(22)$ = -0.16, p = .88, d = 0.03; Second (R^2_{lin} = .98; R^2_{log} = .89; $t(22)$ = -6.94, p < .001, d = 1.40) and third graders (R^2_{lin} = .98; R^2_{log} = .85; $t(22)$ = -5.01, p < .001, d = 1.01) were better fit by the linear function.

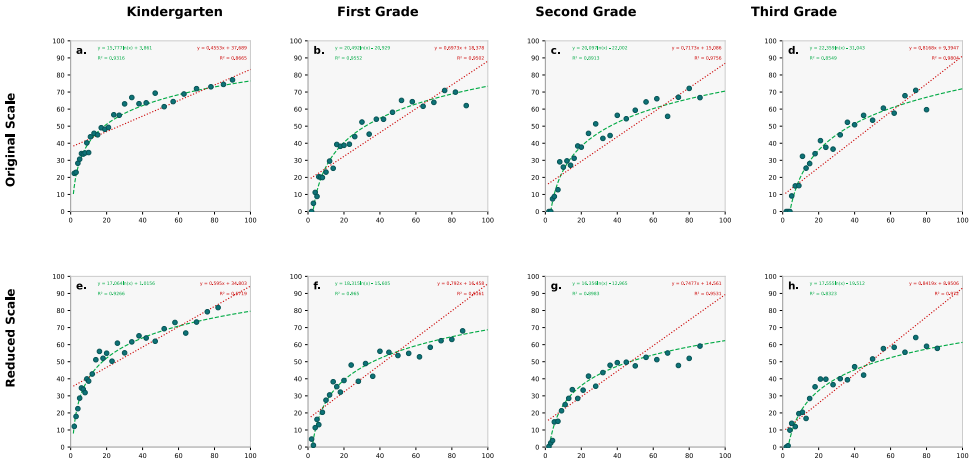
For the reduced scale with 15 items, as shown in Figure 3e, kindergarten students' estimates were better fit by the logarithmic (R^2 = .93) than the linear function (R^2 = .57; $t(14)$ = 3.91, p < .01, d = 0.95). Third graders' estimates on the reduced scale (Figure 3h) were better fit by the linear (R^2 = .97) than the logarithmic function (R^2 = .83; $t(14)$ = -2.99, p < .01, d = 0.73). Estimates on the reduced scale were equally well fit by the linear and logarithmic functions for first (R^2_{lin} = .92; R^2_{log} = .96; $t(14)$ = 1.67, p = .12, d = 0.41) and second graders (R^2_{lin} = .95; R^2_{log} = .90; $t(14)$ = -1.25, p = .23, d = 0.30). That is, findings

1) As in the original studies, the best-fitting exponential function was also computed, but we do not include it here as it was not the best fit for any grade level.

for the reduced scale deviated from the original scale only for second graders with no significant difference in fit between linear and logarithmic with the reduced scale but a better fit with linear for the original scale. These differences will be considered in the discussion. We then proceeded to examine fit distributions at the student level, described next.

Figure 3

Best-Fitting Linear and Logarithmic Functions for Medians at Each Grade for the Original (a-d) and Reduced Scales (e-h)

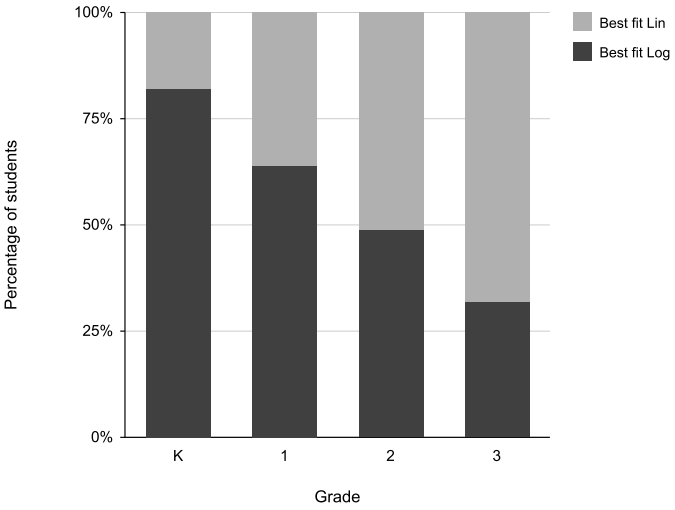


Distribution of the Best-Fitting Type of Function for Individual Students

In the original studies, chi-squared analyses of the distribution of children in each grade for whom the linear or logarithmic function provided the better fit revealed that the percentage of children whose estimates were best fit by the logarithmic function decreased with grade, and the percentage of children best fit by the linear function increased with grade. In general, a higher proportion of kindergartners were best fit by the logarithmic function, a higher proportion of the older students (2nd and/or 3rd graders) were better fit by the linear function, and students in the middle were equally likely to be best fit by the linear and logarithmic functions. With the reduced scale, the best fitting function for students' estimates again varied by grade, $\chi^2(3, N = 234) = 24.408$, $p < .001$. As shown in Figure 4, the percentage of children best fit by the logarithmic function decreased with age, while the percent best fit by the linear function increased.

Figure 4

Best Fit-Measures (Linear vs Logarithmic)



Kindergartners' and first graders' estimates were more likely to be best fit by the logarithmic than the linear function. Second graders' estimates were equally likely to be best fit by the linear and logarithmic functions, while third graders were more likely to be best fit by the linear function than the logarithmic function. Booth and Siegler (2006) also conducted additional paired-sample *t*-tests comparing the mean linear and logarithmic fit at each grade level, revealing that the linear function fit student estimates worse than the logarithmic one for kindergartners, equal to the logarithmic one for first graders, and better than the logarithmic one for second and third graders. The same analyses with the reduced scale across all of the participants revealed that the linear function fit student estimates worse than the logarithmic one for kindergartners [.28 vs. .44; $t(60) = -8.544$, $p < .001$] and first graders [.64 vs. .70; $t(75) = -2.816$ $p < .01$], and better than the logarithmic one for third graders [.84 vs. .76; $t(21) = 2.742$, $p < .01$]. The fit of the linear and logarithmic functions for second graders was equal, with a non-significant trend towards a better fit for the linear function [.72 vs. .69; $t(74) = 1.726$, $p = .09$]. Thus, findings for the reduced scale matched those for the original scale for kindergartners and third graders; for first graders, the reduced scale led to more students being classified as logarithmic than for the original scale, and for second graders the reduced scale led to more students being equally fit by the logarithmic and linear functions rather than better fit by the linear function as in the original scale.

A final way to evaluate the ability of the reduced scale to replicate findings from the original scale is to examine change in the function of best fit for individual students. As shown in Table 6, the best-fitting function from the reduced scale matched that for the original scale for 85.5% of individual students (93.4% of kindergartners, 81.6% of first graders, 84% of second graders, and 81.8% of third graders). Across all grades, when the best-fitting function did not match, the change was more likely to be towards a better fit by the logarithmic function; that is, students who were best fit by the linear function on the original scale but by the logarithmic function for the reduced scale (N = 30 students). Only a handful of students (N = 4) were better fit by the logarithmic function for the original scale but by the linear function for the reduced scale. In general, the reduced scale most accurately replicated the best-fitting function for individual students for kindergartners and was least accurate for second graders.

Table 6

Match in Best-Fitting Function for Individual Students With Original vs. Reduced Scales

Grade	Best fit on Original Scale	Best fit on Reduced Scale		% Match
		Log	Lin	
Kindergarten	Log	48	2	93.4%
	Lin	2	9	
First Grade	Log	36	2	81.6%
	Lin	12	26	
Second Grade	Log	24	0	84.0%
	Lin	12	39	
Third Grade	Log	3	0	81.8%
	Lin	4	15	

Variance Accounted for by the Best Fitting Linear Function (R^2_{lin} Scores) for Individual Students

As in the original studies, we first conducted a one-way ANOVA on R^2_{lin} scores by grade. In the original publications, there were significant main effects of grade on R^2_{lin} scores such that kindergartners' estimates were less linear than those of first or second graders (Siegler & Booth, 2004) or first, second, and third graders (Booth & Siegler, 2006). With the reduced scale, there was also a significant main effect of grade, $F(3,230) = 60.05$, $p < .001$, $\eta^2_p = .439$. Follow-up paired-sample t-tests with Bonferroni correction showed that kindergarteners had lower R^2_{lin} (M = .28) than first (M = .64), second (M = .72), or third graders (M = .85); first graders also had significantly lower R^2_{lin} than third graders (all $p < .001$).

Variance Accounted for by the Mixed Log-Linear Model for Individual Students

We used the method employed by Qin and colleagues (2024) to compute the fit of the mixed log-linear model (MLLM) for each student's estimates, separately for the original scale and for the reduced scale. These calculations resulted in two variables of import for each student for each scale: the amount of variance accounted for by the function (R^2_{MLLM}), and the weight of the logarithmic component of the function (λ). For the data in the present study, R^2_{MLLM} scores did not differ significantly between the original (M = .59) and reduced scales (M = .52; $t(466)$ = .407, p = .68). There was also no significant difference in λ for the original (M = -.09) vs. reduced scales (M = .04; $t(466)$ = -0.677, p = .50). Developmental change in patterns of estimates in Qin and colleagues (2024) was represented by a decrease in λ by student age, indicating that the relative degree of logarithmicity in estimates decreased as students aged. Similarly, in the present study, there were significant decreases in λ by grade for both the original ($F(3,230)$ = 32.82, p < .001, η^2_p = .296) and reduced scales ($F(3,230)$ = 32.25, p < .001, η^2_p = .300). Follow-up paired-sample t-tests with Bonferroni correction for each scale showed that kindergarteners had higher λ (Ms = .91 and .94 for original and reduced scales, respectively) than first (Ms = .59 and .61), second (Ms = .48 and .50), or third graders (Ms = .34 and .35; all ps < .001); first graders also had significantly higher λ than third graders (ps < .01). Descriptive statistics that demonstrate PAE, R^2_{lin} and R^2_{log} by grade are displayed in Table 7.

Table 7

Descriptive Statistics for Original and Reduced Scale by Grade Level

	PAE				R^2_{lin}				R^2_{log}			
	Original Scale		Reduced Scale		Original Scale		Reduced Scale		Original Scale		Reduced Scale	
Grade	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Kindergarten	24.20%	0.052	28.42%	0.081	0.334	0.227	0.284	0.207	0.449	0.248	0.442	0.247
First Grade	15.57%	0.057	16.30%	0.094	0.689	0.235	0.643	0.229	0.714	0.182	0.699	0.190
Second Grade	12.65%	0.045	12.94%	0.075	0.786	0.195	0.721	0.239	0.737	0.159	0.689	0.192
Third Grade	8.88%	0.027	8.93%	0.044	0.879	0.125	0.847	0.115	0.796	0.089	0.758	0.127

Correlations Between Accuracy and Linearity of Estimates With Students' Mathematics Achievement Scores

Because the original studies were conducted at different times and in different school districts, they also took different achievement tests or versions of achievement tests. We therefore first z-standardized achievement test scores within grade within study to ensure that we could compare across studies and across grade levels.

As in the original studies, we computed partial correlations between students' PAE and R^2_{lin} scores with their achievement test scores, controlling for age. Across the origi-

nal studies, there was evidence of connections between both measures of number line performance and student achievement scores, though the particular grade levels and measures that yielded significant correlations varied by study. In general, though, the larger the amount of variance in estimates explained by a linear function (e.g., higher R^2_{lin} scores), and the greater the accuracy of estimates (e.g., lower PAE), the higher the achievement scores. Partial correlations for *z*-standardized achievement test scores with PAE and R^2_{lin} on the reduced scale, controlling for age, can be found in Table 8. In the original studies, all such analyses were conducted separately for each grade level; because we have *z*-standardized the scores to account for the use of different tests that are not identically normed, here we also present the correlations for the whole sample. As with the original papers, with the reduced scale, R^2_{lin} scores were consistently correlated with achievement test scores, such that higher R^2_{lin} scores were associated with higher achievement test scores; this was the case for the sample as a whole as well as for students within each individual grade level. Again, as in the original papers, PAE scores for the reduced scale were correlated with achievement test scores; this was the case across the whole sample, and for first and second graders in particular.

Table 8

Partial Correlations Between Accuracy and Linearity of Number Line Estimates With Achievement Scores, Controlling for Age Using the Reduced Scale

Sample	PAE	R^2_{lin}
All Grades	$r(227) = -.283$, $p < .001$	$r(227) = .318$, $p < .001$
Kindergarten	$r(58) = -.154$, *ns*	$r(58) = .264$, $p < .05$
First Grade	$r(72) = -.411$, $p < .001$	$r(72) = .406$, $p < .001$
Second Grade	$r(71) = -.334$, $p < .01$	$r(71) = .315$, $p < .01$
Third Grade	$r(17) = -.437$, *ns*	$r(17) = .511$, $p < .05$

Discussion

Over the past two decades, the number line estimation task has been widely used as a measure of children's numerical competence (e.g., Li et al., 2024; Ruiz et al., 2023). However, the number of trials given within the task has varied considerably, involving up to 44 or even more trials given to individual students. The present study used data from two previously published studies involving number line estimation on a 0 – 100 scale (Booth & Siegler, 2006; Siegler & Booth, 2004) to determine potential reduced sets of trials that could be used to produce similar results while taking up less time in data collection.

The current findings demonstrate that our reduced whole number line (0 – 100) scale successfully replicates the patterns of developmental findings from both prior

foundational work (Booth & Siegler, 2006; Siegler & Booth, 2004) and recent assertions (Qin et al., 2024). However, we find inconsistencies specifically in relation to patterns of results for second graders, discussed further below. That is, the reduced scale (15 target items) has comparable psychometric properties compared to the original scale (23 target items). Further, as in seminal pieces by Siegler and Booth (Booth & Siegler, 2006; Siegler & Booth, 2004), the logarithmic-to-linear shift is supported with the reduced scale, with kindergarteners' estimates better fit by a logarithmic than a linear function and third graders' estimates better fit by a linear than a logarithmic function. Analysis of individual student distributions also supports this claim. The previous work demonstrated developmental change in how children represent numerical magnitudes on a mental number line. They found that young children (around 5) hold a logarithmic representation of numbers where smaller numbers (e.g., 1, 2, 3) are spaced farther apart on their mental number line, while larger numbers (e.g., 30, 40, 50) are compressed closer together. That is, they overestimate the size of smaller numbers relative to larger numbers. This reflects a non-linear representation of numerical magnitude in which equal intervals represent exponential rather than additive changes. Children's numerical representations become more linear as they develop (~6 – 8 years old). That is, they shift towards holding a linear mental number line in which equal intervals represent the same magnitude as demonstrated by their placement of target numbers spaced equally and appropriately on a number line; a similar finding from Qin and colleagues (2024) regarding increased linearity in student estimates with age is also replicated with our reduced scale.

As in the prior work, first-graders were no more linear in their estimates than logarithmic. However, there were some slight deviations in the findings of the current work and prior work with second graders' estimates. In the seminal work, second graders' estimates were significantly more linear than logarithmic. In the current analysis with the reduced scale, the second graders' estimates did not significantly differ in terms of their linearity or logarithmic nature. An inspection of the R^2 values shows that second graders' estimates seemed to be more linear (R^2_{lin} = .95) than logarithmic (R^2_{log} = .90), but these differences are not significantly different. This discrepancy does not appear to be an issue with power, as the analysis included second graders across three original studies. However, it is possible that the sensitivity of the reduced scale for detecting pattern shifts in number line estimation varies by grade. That is, our reduced scale may be most accurate with respect to linear and logarithmic patterns for the grades chronologically furthest from the developmental transition (kindergartners and third graders) and the grades around the transition (e.g., 2nd grade) may need a more expanded measure to more precisely and fully capture the nuanced change from logarithmic to linear representations.

Next we consider the finding that the best option of the reduced 0 – 100 scales comprised 15 target items, with 12 of those items in the first half of the range of the scale. That is, the scale included the following targets: 3, 4, 6, 8, 12, 17, 21, 24, 25, 29,

33, 39, 52, 57, and 81. This is likely due to the composition of the sample in relation to their developmental stage of numerical magnitude understanding. That is, the majority of the sample held a logarithmic representation of numerical magnitude, and so more target items in the first half of the scale is likely better able to capture the shape of the curve in that part of the scale. As mentioned previously, it is likely that the 0 – 100 scale employed in the present study is not the optimal way to capture the numerical competence of the older students in this study (second and third grade students), and that their performance on a 0 – 1000 scale may be more useful. However, when it is desirable to assess the 0 – 100 number line estimation skills of these older students who have more accurate or linear representations of the scale, researchers may need even fewer target items to capture their magnitude understanding in this range. Future work with older samples may wish to investigate further reductions. Following this logic, it is possible that reduction procedures for larger scales (e.g., 0 – 1000, 0 – 10,000) may follow the same patterns with older children. For example, the logarithmic to linear shift that occurs within the 0 – 1,000 scale happens at the age of 8 – 10 (Booth & Siegler, 2006; Siegler & Booth, 2004). The logarithmic to linear shift that occurs within the 0 – 10,000 scale happens around the age of 10 – 12 (Siegler et al., 2011). It is possible that reduced scales for these ranges would rely more heavily on targets in the 0 – 500 range (for the 0 – 1000 scale) and in the 0 – 5000 range (for the 0 – 10000 scale) if studying 8 – 10 year olds and 10 – 12 year olds, respectively. Alternatively, it is possible that there is already bias reflected in the design of the original scale due to the intentional oversampling of the first half of the scale. That is, because there are more target items in the first half of the number line, it may be easier to replicate a logarithmic curve found with a reduced set of target items sampled from this same distribution. To determine whether this possibility is a better explanation for the perceived accuracy of the reduced scale, a set of targets that are equally representative of the entirety of the scale may be necessary. Though this is a possibility, and more research with a wider variety of targets may serve to support or challenge prior work, we consider the findings of Gashaj et al. (2016), who used an approximately equal number of target items above and below 50 on a 0 – 100 scale and found that kindergarteners' numerical magnitude on a 0 – 100 number line was more logarithmic than linear. Still, further research using a more even distribution of target items across scales at different grades or ages may clarify whether potential biases exist in the original scale.

We also employed rigorous psychometric methods to evaluate and refine the scale, addressing limitations in the original scale's development. We utilized the graded response model (GRM), which is specifically designed for analyzing ordered polytomous data such as Likert-style items (Samejima, 2016). This approach allowed us to examine item-level properties and overall scale performance with greater precision than classical test theory methods alone.

Our analyses also included exploratory factor analysis (EFA) with scree plot examination, providing insights into the scale's dimensionality (Cattell, 1966). We also conducted parallel analysis, which is considered one of the most accurate methods for determining factor retention (Horn, 1965). The use of multiple criteria for assessing dimensionality, including eigenvalue ratios and explained variance, strengthened our conclusions about the scale's structure. We also used both Cronbach's alpha and McDonald's omega to check for reliability, which gave us a more complete picture of internal consistency (Dunn et al., 2014). These methodological strengths, combined with our systematic approach to scale reduction, provide a solid foundation for confidence in our findings. The refined scale not only maintains the psychometric integrity of the original but also potentially improves its unidimensionality and efficiency, addressing the limitations of the original scale's atheoretical development.

Despite these strengths, the current study has the following limitations that can be addressed in future research. First, we note some methodological limitations. Our method for factor analysis was exploratory in nature. In future studies, confirmatory factor analysis (CFA) could be used to confirm the structure of the scale even more. It would be possible to test the hypothesized factor structure that came out of our exploratory analyses more thoroughly with CFA. This could make the construct validity of the refined scale stronger. Moreover, the assumptions of unidimensionality and local independence, which are fundamental to the graded response model used in this study, warrant further examination. Future research should investigate how well these assumptions hold across different contexts and populations, as violations of these assumptions could impact the accuracy of item parameter estimates and overall scale performance.

Further, while we believe that the reduced scale sufficiently captures variance in students' number line estimation skills, it is, of course, not certain that they would have made the same estimations on the target items if they had only been given those items. That is, it is possible that the students' estimations on the retained target items were influenced by their experience considering the other, non-retained target items. Future work would need to directly compare children's estimations when given only the reduced scale vs. the full scale to determine whether this is of consequence on estimates. Additionally, while our sample (N = 234) was drawn from three studies by the same group—which may limit external generalizability—we note that the present work primarily targets scale reduction and item functioning (discrimination, thresholds) rather than population-level generalization; nonetheless, future validation in independent, diverse samples is warranted. We also note here that we had far fewer third graders in our analytic sample than earlier grades. Future work in this area should attempt to recruit a more even distribution of participants per grade.

We also consider the age of the data itself. That is, the data used in the current study originates from studies conducted as early as 2004. Some of the trends we see in the logarithmic to linear shift may happen at a different age or grade level, as schools

have taken steps to focus more on inclusion of number line activities in their standard mathematical practices in second grade (e.g., National Governors Association Center for Best Practices & Council of Chief State School Officers [NGA & CCSSO], 2010). More current number line work with the 0 – 100 scale has replicated the same pattern but has argued that the pattern is more reflective of skill at mapping numbers to space (e.g., Cohen & Quinlan, 2018; Haman & Patro, 2022; Sasanguie et al., 2016). Indeed, most recent work using the 0 – 100 NLE task emphasizes results regarding accuracy (e.g., PAE) rather than linearity (e.g., Ruiz et al., 2024); or evaluate linearity without consideration of the logarithmic function (e.g., Li et al., 2024). Thus, the usefulness of the NLE task for comparing patterns of estimation may be less important now. Researchers who maintain the goal of investigating transitions in estimation patterns may need to confirm that the reduced scale we derived in the current study yields the same or similar findings with children currently in third grade or below.

It is also important to note that although discrimination might differ by grade, our per-grade subsamples are too small to support stable multi-group GRM estimation; small group sizes in polytomous IRT can lead to imprecise discrimination and threshold estimates, so we report a single-group calibration and reserve formal grade-level invariance tests for future work with larger, grade-balanced samples.

Finally, it is worth noting that the reduced scale is comprised almost entirely of items in the lower half of the numerical scale. When the original task was designed, numbers were oversampled from the lower third of the scale in order to capture a potential logarithmic-to-linear shift (e.g., Siegler & Booth, 2004). Oversampling in the case of the 0 – 100 number line meant including 10 items between 0 – 30 and 14 items between 31 – 100. In contrast, our reduced scale retained 8 items between 0 – 30 and only 5 items between 31 – 100. Thus, oversampling of the low end of the scale is indeed important and may in fact be even more critical than originally anticipated. That is, accurately capturing variance in young students' number line estimations appears to require an even greater proportion of numbers at the low end and many fewer items in the top two-thirds of the scale.

Practical Implications

We consider the practical implications of this work. Magnitude understanding is one of the most studied mathematical competencies in the fields of both psychology and education. Thus, the number line task is one of the most often administered assessments in research in this domain. Though some researchers use a paper-and-pencil task as was traditionally done in the foundational work (e.g., Booth & Siegler, 2006), much of the field is moving towards iPad or Chromebook administration of the task. Administration is typically conducted in a one-on-one setting with a researcher. Math instructional time makes up approximately 12% of U.S. students' school days (with variations across states and districts; Mullis et al., 2016). Many schools are hesitant to interrupt children's

school day to engage in research. U.S. schools in particular are hesitant to spend too much time on research-based activities that they see as taking away from instructional time. Although intervention-based work can be seen as instructional time and may be welcome by many school administrators and practitioners, the assessments that come along with such classroom-based research are typically not viewed as "instructional practices". That is, schools prioritize instructional time, and research activities must not interfere with teaching and learning. Thus, schools actively safeguard instructional time. As such, researchers conducting classroom-based assessments must be cognizant of the time their assessments take. The current scale reduction study takes important strides towards addressing these concerns in both a practical and methodologically rigorous way that simultaneously aligns with the developmental changes that occur within the children the scale is designed to assess.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Previously Presented: An earlier version of this work was presented at the 2025 AERA Annual Meeting. The conference presentation introduced the study concept and preliminary results whereas the present article reports the full analyses, complete results, and expanded implications.

Data Availability: The data supporting the findings of this study are not publicly available because they are part of an ongoing larger longitudinal research project. To support transparency and reproducibility, metadata and analytic materials (e.g., study documentation and R code) are available from the corresponding author upon reasonable request.

References

Ashkenazi, S., & Cohen, N. (2023). Developmental trajectories of number line estimations in math anxiety: Evidence from bounded and unbounded number line estimation. *Applied Cognitive Psychology, 37*(6), 1316–1327. https://doi.org/10.1002/acp.4125

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Barbieri, C. A., Miller-Cotto, D., Clerjuste, S. N., & Chawla, K. (2023). A meta-analysis of the worked examples effect on mathematics performance. *Educational Psychology Review, 35*(1), Article 11. https://doi.org/10.1007/s10648-023-09745-1

Barbieri, C. A., Young, L. K., Newton, K. J., & Booth, J. L. (2021). Predicting middle school profiles of algebra performance using fraction knowledge. *Child Development, 92*(5), 1984–2005. https://doi.org/10.1111/cdev.13568

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

Berteletti, I., Man, G., & Booth, J. R. (2015). How number line estimation skills relate to neural activations in single digit subtraction problems. *NeuroImage, 107*, 198–206. https://doi.org/10.1016/j.neuroimage.2014.12.011

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology, 37*(4), 247–253. https://doi.org/10.1016/j.cedpsych.2012.07.001

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of Experimental Child Psychology, 118*, 110–118. https://doi.org/10.1016/j.jecp.2013.09.001

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

Booth, J. L., & Siegler, R. S. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development, 79*(4), 1016–1031. https://doi.org/10.1111/j.1467-8624.2008.01173.x

Brez, C. C., Miller, A. D., & Ramirez, E. M. (2016). Numerical estimation in children for both positive and negative numbers. *Journal of Cognition and Development, 17*(2), 341–358. https://doi.org/10.1080/15248372.2015.1033525

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chan, J. Y.-C., & Mazzocco, M. M. (2024). New measures of number line estimation performance reveal children's ordinal understanding of numbers. *Journal of Experimental Child Psychology, 245*, Article 105965. https://doi.org/10.1016/j.jecp.2024.105965

Cohen, D. J., & Quinlan, P. T. (2018). The log–linear response function of the bounded number-line task is unrelated to the psychological representation of quantity. *Psychonomic Bulletin & Review, 25*(1), 447–454. https://doi.org/10.3758/s13423-017-1290-z

Cornu, V., Hornung, C., Schiltz, C., & Martin, R. (2017). How do different aspects of spatial skills relate to early arithmetic and number line estimation? *Journal of Numerical Cognition, 3*(2), 309–343. https://doi.org/10.5964/jnc.v3i2.36

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*(5880), 1217–1220. https://doi.org/10.1126/science.1156540

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). From rational numbers to algebra: Separable contributions of decimal magnitude and relational understanding of fractions. *Journal of Experimental Child Psychology, 133*, 72–84. https://doi.org/10.1016/j.jecp.2015.01.013

Dietrich, J. F., Huber, S., Dackermann, T., Moeller, K., & Fischer, U. (2016). Place-value understanding in number line estimation predicts future arithmetic performance. *The British Journal of Developmental Psychology, 34*(4), 502–517. https://doi.org/10.1111/bjdp.12146

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Ellis, A., Susperreguy, M. I., Purpura, D. J., & Davis-Kean, P. E. (2021). Conceptual replication and extension of the relation between the number line estimation task and mathematical competence across seven studies. *Journal of Numerical Cognition, 7*(3), 435–452. https://doi.org/10.5964/jnc.7033

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Psychology Press.

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. https://doi.org/10.1016/j.jecp.2014.01.013

Fitzsimmons, C. J., & Thompson, C. A. (2022). Developmental differences in monitoring accuracy and cue use when estimating whole-number and fraction magnitudes. *Cognitive Development, 61*, Article 101148. https://doi.org/10.1016/j.cogdev.2021.101148

Gashaj, V., Uehlinger, Y., & Roebers, C. M. (2016). Numerical magnitude skills in 6-years-old children: Exploring specific associations with components of executive function. *Journal of Educational and Developmental Psychology, 6*(1), 157–172. https://doi.org/10.5539/jedp.v6n1p157

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology, 104*(1), 206–223. https://doi.org/10.1037/a0025398

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

Gunderson, E. A., & Hildebrand, L. (2021). Relations among spatial skills, number line estimation, and exact and approximate calculation in young children. *Journal of Experimental Child Psychology, 212*, Article 105251. https://doi.org/10.1016/j.jecp.2021.105251

Haman, M., & Patro, K. (2022). More linear than log? Non-symbolic number-line estimation in 3-to 5-year-old children. *Frontiers in Psychology, 13*, Article 1003696. https://doi.org/10.3389/fpsyg.2022.1003696

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* SAGE.

Hoard, M. K., Geary, D. C., Byrd-Craven, J., & Nugent, L. (2008). Mathematical cognition in intellectually precocious first graders. *Developmental Neuropsychology, 33*(3), 251–276. https://doi.org/10.1080/87565640801982338

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Jung, S., Roesch, S., Klein, E., Dackermann, T., Heller, J., & Moeller, K. (2020). The strategy matters: Bounded and unbounded number line estimation in secondary school children. *Cognitive Development, 53*, Article 100839. https://doi.org/10.1016/j.cogdev.2019.100839

Landy, D., Charlesworth, A., & Ottmar, E. (2017). Categories of large numbers in line estimation. *Cognitive Science, 41*(2), 326–353. https://doi.org/10.1111/cogs.12342

Li, M., Yang, J., & Ye, X. (2024). Children's number line estimation strategies: Evidence from bounded and unbounded number line estimation tasks. *Frontiers in Psychology, 15*, Article 1421821. https://doi.org/10.3389/fpsyg.2024.1421821

Link, T., Huber, S., Nuerk, H.-C., & Moeller, K. (2014). Unbounding the mental number line—New evidence on children's spatial representation of numbers. *Frontiers in Psychology, 4*, Article 1021. https://doi.org/10.3389/fpsyg.2013.01021

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center.

Namkung, J. M., & Fuchs, L. S. (2016). Cognitive predictors of calculations and number line estimation with whole numbers and fractions among at-risk students. *Journal of Educational Psychology, 108*(2), 214–228. https://doi.org/10.1037/edu0000055

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. https://corestandards.org

Nuraydin, S., Stricker, J., Ugen, S., Martin, R., & Schneider, M. (2023). The number line estimation task is a valid tool for assessing mathematical achievement: A population level study with 6484 Luxembourgish ninth-graders. *Journal of Experimental Child Psychology, 225*, Article 105521. https://doi.org/10.1016/j.jecp.2022.105521

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*, Article 1082. https://doi.org/10.3389/fpsyg.2017.01082

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

Qin, J., Kim, D., & Opfer, J. E. (2024). Varieties of number-line estimation: Systematic review, models, and data. *Developmental Review, 74*, Article 101161. https://doi.org/10.1016/j.dr.2024.101161

Reinert, R. M., Hartmann, M., Huber, S., & Moeller, K. (2019). Unbounded number line estimation as a measure of numerical estimation. *PLoS One, 14*(3), Article e0213102. https://doi.org/10.1371/journal.pone.0213102

Revelle, W. (2021). *psych: procedures for psychological, psychometric, and personality research* (R package version 1.9.12) [Computer software]. Northwestern University, Evanston, Illinois. https://cran.r-project.org/web/packages/ltm/index.html

Rizopoulos, M. D. (2018). *Package 'ltm'* [Computer software]. http://wiki.r-project.org/rwiki/doku.php

Ruiz, C., Kohnen, S., & Bull, R. (2023). Number line estimation patterns and their relationship with mathematical performance. *Journal of Numerical Cognition, 9*(2), 285–301. https://doi.org/10.5964/jnc.10557

Ruiz, C., Kohnen, S., & Bull, R. (2024). The relationship between number line estimation and mathematical reasoning: A quantile regression approach. *European Journal of Psychology of Education, 39*(2), 581–606. https://doi.org/10.1007/s10212-023-00708-2

Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 95-107). Chapman and Hall/CRC.

Sarı, M. H., & Olkun, S. (2021). Number line estimations, place value understanding and mathematics achievement. *Journal of Education and Future, 19*, 37–47. https://doi.org/10.30786/jef.729843

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*(4), 382–405. https://doi.org/10.5334/pb.276

Schiller, L. K., Abreu-Mendoza, R. A., Thompson, C. A., & Rosenberg-Lee, M. (2024). Children's estimates of equivalent rational number magnitudes are not equal: Evidence from whole numbers, percentages, decimals, and fractions. *Journal of Experimental Child Psychology, 247*, Article 106030. https://doi.org/10.1016/j.jecp.2024.106030

Schneider, M., Grabner, R. H., & Paetsch, J. (2009). Mental number line, number line estimation, and mathematical achievement: Their interrelations in Grades 5 and 6. *Journal of Educational Psychology, 101*(2), 359–372. https://doi.org/10.1037/a0013840

Schneider, M., Merz, S., Stricker, J., De Smedt, B., 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., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Silla, E. M., Guba, T. P., Rodrigues, A., Anisiobi, O. C., Scanniello, A., & Barbieri, C. A. A. (2026). *Systematic review of mathematical and motivational relations with algebra performance* [Manuscript under review].

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102*, 443–461. https://doi.org/10.1007/s11205-010-9682-8

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General, 142*(1), 193–208. https://doi.org/10.1037/a0028560

Steinke, D. A. (2017). Evaluating number sense in community college developmental math students. *COABE Journal, 6*(1), 5–19.

Sudo, S., Chileya, G., Kume, A., & Fujino, Y. (2022). Estimating number line as a cause of low mathematics performance in Zambia. In *Proceedings of the 7th International STEM Education Conference 2022 (iSTEM-Ed), Sukhothai, Thailand.* IEEE.

Sullivan, J. L., Juhasz, B. J., Slattery, T. J., & Barth, H. C. (2011). Adults' number-line estimation strategies: Evidence from eye movements. *Psychonomic Bulletin & Review, 18*(3), 557–563. https://doi.org/10.3758/s13423-011-0081-1

Tremolada, M., Taverna, L., Bonichini, S., Pillon, M., & Biffi, A. (2019). The developmental pathways of preschool children with acute lymphoblastic leukemia: Communicative and social sequelae one year after treatment. *Children, 6*(8), Article 92. https://doi.org/10.3390/children6080092

Vogel, S. E., Grabner, R. H., Schneider, M., Siegler, R. S., & Ansari, D. (2013). Overlapping and distinct brain regions involved in estimating the spatial position of numerical and non-numerical magnitudes: An fMRI study. *Neuropsychologia, 51*(5), 979–989. https://doi.org/10.1016/j.neuropsychologia.2013.02.001

Wall, J. L., Thompson, C. A., Dunlosky, J., & Merriman, W. E. (2016). Children can accurately monitor and control their number-line estimation performance. *Developmental Psychology, 52*(10), 1493–1502. https://doi.org/10.1037/dev0000180

Xu, C., Burr, S. D. L., LeFevre, J.-A., Skwarchuk, S.-L., Osana, H. P., Maloney, E. A., Wylie, J., Simms, V., Susperreguy, M. I., Douglas, H., & Lafay, A. (2023). Development of children's number line estimation in primary school: Regional and curricular influences. *Cognitive Development, 67*, Article 101355. https://doi.org/10.1016/j.cogdev.2023.101355

Young, L. K., & Booth, J. L. (2015). Student magnitude knowledge of negative numbers. *Journal of Numerical Cognition, 1*(1), 38–55. https://doi.org/10.5964/jnc.v1i1.7

Yu, S., Kim, D., Fitzsimmons, C. J., Mielicki, M. K., Thompson, C. A., & Opfer, J. E. (2022). From integers to fractions: The role of analogy in developing a coherent understanding of proportional magnitude. *Developmental Psychology, 58*(10), 1912–1930. https://doi.org/10.1037/dev0001398

Zhu, M., Cai, D., & Leung, A. W. (2017). Number line estimation predicts mathematical skills: Difference in Grades 2 and 4. *Frontiers in Psychology, 8*, Article 1576. https://doi.org/10.3389/fpsyg.2017.01576

MCLS THE MATHEMATICAL COGNITION AND LEARNING SOCIETY

Journal of Numerical Cognition (JNC) is the official journal of the Mathematical Cognition and Learning Society (MCLS).

zpid Leibniz-Institut für Psychologie

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.