# Investigating the Utility of a Kindergarten Number Line Assessment Compared to an Early Numeracy Screening Battery

Marah Sutherland[a, b], Ben Clarke[a, b], Joseph F. T. Nese[b], Mari Strand Cary[a, b], Lina Shanley[a, b], David Furjanic[a], & Lillian Durán[a, b]

[a]Center on Teaching and Learning
1600 Millrace Drive, Suite 207
Eugene, OR, 97403
USA

[b]University of Oregon
1585 E 13th Ave.
Eugene, OR, 97403
USA

**Corresponding Author:**

Marah Sutherland
Center on Teaching and Learning
1600 Millrace Drive, Suite 207
Eugene, OR, 97403
marahs@uoregon.edu

Received 15 May 2020, Revised 21 August 2020, Accepted 9 November 2020, Available online 5 December 2020.
Sutherland, M., Clarke, B., Nese, J. F. T., Strand Cary, M., Shanley, L., Furjanic, D., & Durán, L. (2021). Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery. *Early Childhood Research Quarterly, 55,* 119–128. https://doi.org/https://doi.org/10.1016/j.ecresq.2020.11.003

Funding: **R324A110286** (Institute of Education Sciences, NCSER) & **H327S140019** (Office of Special Education Programs, Office of Special Education and Rehabilitative Services)

**Declaration of Interest:** The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Mari Strand Cary, Lina Shanley, and Ben Clarke are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain KinderTEK-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services.

**Funding:** The research reported here was supported by the U.S. Department of Education through the Institute of Education Sciences Grant R324A110286 and the Office of Special Education Programs Grant H327S140019, both to the University of Oregon. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred.

**Acknowledgements:** Appreciation is expressed to educators and students who participated in the KinderTEK project and to the KinderTEK team, particularly Marcia Moore and Cathy Watkins, for ongoing assistance and support.

# Investigating the Utility of a Kindergarten Number Line Assessment Compared to an Early Numeracy Screening Battery

## Abstract

Drawing from the developmental and cognitive mathematics literature, the purpose of this study was to investigate the reliability, validity, and diagnostic utility of a widely-researched number line task in kindergarten. Specifically, the Number Line Assessment 0-100 (NLA 0-100) as compared to an established kindergarten screening measure was examined using (a) regression models and (b) classification accuracy. Five- and six-year-old students (*n* = 154) were assessed on numeracy measures in the fall and spring of their kindergarten year. The NLA 0-100 had lower predictive validity and lower classification accuracy compared to an early numeracy screening measure. The findings have implications for practice and future research using number line screening assessments.

Keywords: mathematics, screening, number line estimation task, early numeracy

# Introduction

Prior to the start of formal schooling, children have varying amounts of exposure to and experience with mathematics. Parents differ in the extent to which they engage their child in “number talk” at home, promote counting practice or numeral identification in everyday life, and provide mathematics support in the home environment (Anders et al., 2012; Clements et al., 2003). These types of early mathematics experiences are related to the development of informal number knowledge and associated with success upon kindergarten entry (LeFevre et al., 2009; Ramani & Siegler, 2008). Disparities also occur in students’ access to high-quality preschool programs that explicitly teach and focus on mathematical concepts (Anders et al., 2012; Melhuish et al., 2008). Early differences in exposure to number across multiple settings result in mathematics opportunity gaps for students as young as preschool-age, with students from upper- and middle-class backgrounds already achieving higher than their economically disadvantaged peers (Saxe et al., 1987; Starkey et al., 2004). Once established, gaps between high- and low-achieving students tend to persist and widen as students encounter increasingly advanced mathematical content throughout the elementary school grades (Judge & Watson, 2011; Morgan et al., 2009).

As the first exposure to formal schooling, and for many the first formal introduction to mathematics, kindergarten represents a particularly salient time to intervene for struggling students and to alter stable trajectories of poor mathematics performance (Morgan et al., 2009). Within Response to Intervention (RTI) or Multi-Tiered Systems of Support (MTSS) models, schools must be efficient in their allocation of intervention resources, by selecting students for intervention that are at risk. To successfully operate within RTI or MTSS, schools need to make accurate decisions about which students should be selected for intervention. Mathematics screening tools are critical to enable identification of at-risk students early on and provide them with intervention prior to falling further behind their typically-achieving peers (Author et al., 2014; Gersten et al., 2012).

In the current study, we investigated the utility of a number line screening measure as compared to an established set of early numeracy curriculum-based measures (CBMs) to identify kindergarten students at risk for developing mathematics difficulties. We chose kindergarteners as our target sample due to our interest in critical underlying mathematics constructs, such as the number line, that may enable more accurate screening for risk as students transition from informal to formal mathematics. Additionally, accurate screening in kindergarten is of high interest due to the relatively stable trajectories of mathematics performance from kindergarten to the later elementary grades (e.g., Morgan et al., 2009). In the remainder of the Introduction, we describe the role of number sense in informing mathematics screening measures, the mental number line construct and its relation to number sense, and the potential application of a number line assessment as a screening tool for use with kindergarten-aged students.

## The Role of Number Sense in Early Mathematics Development

One type of number knowledge underlying the development of important mathematical skills in kindergarten and beyond is *number sense*. While the exact definition of number sense is multifaceted, researchers generally agree that it signifies an in-depth understanding of number and flexibility in solving mathematical problems (Berch, 2005; Gersten et al., 2005). More detailed descriptions of number sense specify critical skills and mathematical concepts including preverbal mathematical numeration (i.e., recognizing when one item is added to or taken from a set), understanding of symbolic number representations (e.g., the numeral "5" represents five entities), understanding relations among numbers and number magnitudes (e.g., understanding that 6 is one more than 5, and 7 is one more than 6), and use of estimation strategies including awareness of unreasonable results (Griffin et al., 1994; Kalchman et al., 2001; Siegler et al., 2011). Number sense has also been described as a mechanism that students use to organize their understanding of number – such as relying on benchmark numbers, patterns in the base ten system, or conceptual structures such as a mental number line, when solving mathematical problems (Case & Okamoto, 1996; Greeno, 1991; Griffin, 2004). Fourteen of the 22 kindergarten Common Core State Standards in Mathematics (CCSS-M, 2010) can be tied to number sense elements, including numeral and number recognizion, magnitude comparisons, counting principles, fact fluency, and math language (Witzel, Roccomini, & Herlong, 2013). While number sense is often associated with early mathematics development, its impact spans far beyond. Strong number sense sets the foundation for more advanced mathematical skills, such as the acquisition of complex arithmetic procedures and an understanding of number systems (Berch, 2005; Gersten & Chard, 1999).

## Screening for Number Sense

Given the importance of number sense to early and later mathematics development, one commonality across early numeracy screening measures is a focus on assessing components of number sense (Gersten et al., 2012). This has largely been accomplished through the development of discrete tasks that are believed to tap into the essential components of number sense (Mazzocco, 2005). These tasks typically follow CBM design parameters, with characteristics such as a brief administration time, ease of use, and a focus on tasks that are sensitive to student growth (Deno, 1985). Early numeracy CBM measures have included tasks such as comparing number magnitudes, determining the missing number in a sequence, demonstrating understanding of counting principles, and solving arithmetic problems requiring base ten understanding (Chard et al., 2005; Author et al., 2011a; Author et al., 2004; Conoyer et al., 2016; Hampton et al., 2012; Lembke & Foegen, 2009; Seethaler & Fuchs, 2010). For example, using a sample of 436 kindergarten students, Chard et al. (2005) found that the predictive validity of one-minute timed measures, including a measure of number identification, quantity discrimination, and strategic counting, ranged from .53 to .61 with the Number Knowledge Test (Okamoto & Case, 1996) from fall to spring. Hampton et al. (2012) administered a measure of counting, number identification, strategic counting, and quantity discrimination to 71 kindergarten students, and found that predictive validities from fall to spring ranged from .26 to .45 with the Broad Math Score of the Woodcock-Johnson Battery of Achievement-III (Woodcock et al., 2001). Seethaler and Fuchs (2010) used a sample of 196 kindergarten students and administered two multi-skill mathematics screeners, including a group-administered computation fluency measure and a number sense measure, along with a single-skill screener focused on quantity discrimination. The researchers found that the predictive validity of the fall screening measures with spring outcome measures ranged from .34 to .75, depending on the screening and outcome measures examined. The researchers also examined the classification accuracy of the fall screening measures, using the Math Reasoning and Numerical Operations subtests of the Early Math Diagnostic Assessment (The Psychological Corporation, 2002) as the outcome measures administered in the spring of Grade 1. Holding sensitivity constant at .90, they found that specificity ranged from .32 to .66, and the area under the curve (AUC) ranged from .67 to .86, with better values associated with the Math Reasoning subtest as the outcome measure.

These types of screening measures have shown promise for identifying students as at risk in kindergarten, demonstrating moderate predictive validities and classification accuracy.

However, in contrast to screening in other content areas such as early literacy or even compared to screening at later grades in mathematics, the predictive validity and diagnostic utility of kindergarten numeracy screening measures is not as strong. This is particularly problematic given concerns related to false positives at kindergarten entry and the need for schools to judiciously allocate resources (Seethaler & Fuchs, 2010). For example, Lembke and Foegen (2009) administered four one-minute timed measures to a sample of kindergarten and first grade students: quantity discrimination, quantity array, number identification, and missing number. They found that predictive validities from fall to spring with the Test of Early Mathematics Achievement-III (Ginsburg & Baroody, 2003) ranged from .34 to .37 in kindergarten, and .43 to .68 in first grade. The challenge of mathematics screening in kindergarten is also illustrated in a review of early numeracy screening research published between 1996 and 2011 by Gersten et al. (2012). The researchers found that the median predictive validity from fall to spring of quantity discrimination tasks was .62 in first grade, but .50 in kindergarten, with outcome measures largely consisting of mathematics achievement tests and one mathematics computation measure. Similarly, for strategic counting tasks, the median predictive validity was .62 in first grade and .48 in kindergarten, again with mathematics achievement tests primarily serving as outcome measures.

Also contributing to the challenge in developing valid screening measures in kindergarten is the lack of consensus among researchers as to which specific early numeracy skills are most indicative of later mathematics difficulties (Gersten et al., 2005; Mazzocco, 2005). Given that early numeracy skills develop in a less linear manner compared to other academic content areas such as early literacy, different strands of mathematical knowledge may develop separately, with some areas developing more quickly than others (Fuchs et al., 2008). This has typically led to mathematics screening measures adopting a widened scope to include assessment of multiple skills (Seethaler & Fuchs, 2010). Including multiple measures in a screening battery, such as quantity discrimination and strategic counting tasks, has yielded moderate success for screening measures in kindergarten. Yet one approach that is underexplored in mathematics screening is to instead focus on underlying constructs that may unify how students approach solving these types of discrete tasks. This issue is reflected in a call from Mazzocco (2005), to think more critically about tasks that have been demonstrated to be good screeners in early mathematics, and how these tasks are tapping into the development of mathematics skills over time. Given the relative success of magnitude comparison and strategic counting tasks as screeners, an alternative approach to assessing discrete mathematical skills is to investigate central conceptual structures that may underlie these individual skills and reflect how students develop mathematical thinking.

## The Mental Number Line Construct

Researchers theorize that students use a mental number line as a central conceptual structure to solve the types of tasks commonly included in early numeracy screeners, such as comparing number magnitudes (Case & Okamoto, 1996; Laski & Siegler, 2007; Schneider et al., 2018). The mental number line has been extensively described by developmental and cognitive mathematics researchers as an underlying feature of number sense (Case & Okamoto, 1996; Greeno, 1991) which students utilize as they engage with and solve early numeracy problems. For example, when comparing number magnitudes (e.g., 6 and 9; 17 and 12), a student would visualize where the two numbers fall on a number line to determine which is greater or when determining a missing value in a number sequence (e.g., __, 11, 12; 66, 67, __), a student would reference a mental number line to support their understanding of number order. Similarly, when students are solving addition and subtraction problems (e.g., 6 + 2; 8 – 4), they rely on a mental number line as they count up or count down to determine the answer. Due to the central role that the number line plays in solving mathematics tasks and its foundational role in number sense (Case & Okamoto, 1996), directly assessing students' knowledge and use of the number line offers promise for early numeracy screening.

The most widely-researched measure used in the developmental and cognitive mathematics literature to assess number line understanding is the number line estimation task (Berteletti et al., 2010; Geary et al., 2008; Laski & Siegler, 2007; Siegler & Booth, 2004, Siegler & Opfer, 2003). While variations of this task exist regarding modality (e.g., paper and pencil, computer-based) and scoring procedure, most commonly this task involves presenting students with a blank number line with two endpoints (e.g., 0 and 100) and then providing students with target numerals to place along the number line. Typically student responses are scored based on the average absolute distance from their placement of the target numeral along the number line to the true location of the target numeral, sometimes reported as a percentage (e.g., dividing the absolute error by the scale of estimates, such as 100 on a 0-100 number line; e.g., Siegler & Booth, 2004). For example, if a student placed the target numeral "20" in the actual location of "25" along a 0-100 number line, the student would earn an error score of "5" for that item, or 5%. Error scores are averaged across all numerals presented, resulting in a mean absolute error score for each student.

Student patterns of performance on the number line estimation task demonstrate that as students increase in age and mathematical proficiency, their performance on the task improves and they are able to complete increasingly harder number line tasks with greater accuracy (e.g., transitioning from a 0-100 number line to a 0-1,000 number line; Siegler & Booth, 2004; Siegler & Opfer, 2003). Researchers investigating this task have also found that more accurate numeral placement on the number line is predictive of higher overall mathematics achievement, better performance on magnitude comparison tasks, and greater understanding of fractions in the upper elementary grades (Booth & Siegler, 2006; Jordan et al., 2013; Laski & Siegler, 2007; Schneider et al., 2018). Prior research has also found that number line task is predictive of mathematics achievement while controlling for executive functions (Barth & Paladino, 2011), working memory (Geary et al., 2007), and cognitive abilities (Zhu et al., 2017).

A recent meta-analysis conducted by Schneider et al. (2018) reviewed number line studies, including over 10,000 participants and averaging over 263 effect sizes. The researchers found that the number line estimation task is a strong measure for predicting mathematics achievement, across grade levels (K-8), number line content (whole number versus fraction number lines), and methodological variations. Across the studies reviewed, the number line estimation task had stronger correlations with students' mathematical performance ($r$ = .44) compared to tasks used in current mathematics screening batteries, such as the magnitude comparison task ($r$ = .27).

While these studies speak to heightened interest in the number line construct in the developmental and cognitive mathematics literature, to our knowledge, only one exploratory study has specifically focused on investigating the number line estimation task in conjunction with commonly used screening tasks. Authors (2018) investigated whether two versions of the number line estimation task, spanning 0-20 and 0-100, added value to an early numeracy screening battery in predicting student outcomes. Assessments were administered to 46 exiting kindergarteners and 60 exiting first graders. All participating students were administered assessments before and after participation in a five-week summer school program; thus, the sample was lower-performing and pre- and post-assessments were administered over a short period of time. While the 0-20 number line task did not show promise for either grade level, the 0-100 task predicted exiting first graders' percentile rank scores on the easyCBM mathematics progress monitoring assessment (Alanzo et al., 2006), explaining 13% of additional variance above and beyond a standardized early numeracy screening battery. For kindergarten students, the number line estimation task explained 7% of additional variance though this result was not statistically significant. While providing initial evidence for the potential of the number line task to screen for mathematics difficulty, the study had a number of limitations, including the small sample that was lower-performing in nature, and the short period of time between assessments. Additionally, the researchers did not investigate the diagnostic utility of the number line assessment, which would provide useful information to schools regarding the number of students correctly classified as at-risk or not at-risk by the number line task.

## Research Questions

Previous screening research in early mathematics has largely focused on evaluating discrete mathematics skills. Given that students are theorized to use and reference a mental number line when solving the types of tasks commonly assessed in early numeracy screeners, and the strong links found between the number line task and student mathematics achievement in cognitive and developmental psychology fields, more research investigating the number line task as a screener is warranted. The current study builds upon prior number line screening research in three key ways. First, we used a broader sample of students with the full range of mathematics skill typically found in a kindergarten classroom. Second, we administered assessments across the academic year, from the fall to spring of kindergarten. Third, given the decisions that schools must make regarding classification of students as at-risk or not at-risk to guide allocation of intervention supports, we build on previous number line research by providing classification accuracy analyses and directly comparing the number line measure to an established early numeracy measure using the same sample of students. In the current study, we investigated the use of an iPad-based number line assessment (NLA 0-100; Authors et al., 2014) spanning 0 to 100. We compared the number line measure to a set of previously-established early numeracy CBMs, Assessing Student Proficiency in Early Number Sense (ASPENS; Sopris; Author et al., 2011a), in predicting students' mathematics scores from the fall to the spring of kindergarten. To address this objective, we examined regression models and classification accuracy to investigate the utility of the number line measure as a screener compared to existing measures in the field. Our research questions were as follows: (1) What are the associations between the NLA 0-100, the ASPENS, and other measures of early numeracy? (2) What proportion of variance do the NLA 0-100 and the ASPENS individually explain in students' spring mathematics scores on the Number Sense Brief? (NSB; Jordan, Glutting, et al., 2008), and (3) What is the diagnostic utility of the NLA 0-100 and the ASPENS? Given the extent of the research on the number line task in the cognitive and developmental mathematics literature, and the theorized role of the number line as a critical underlying construct, we hypothesized that the NLA 0-100 would show similar associations with other numeracy measures, variance explained, and classification accuracy as the ASPENS.

## Material and Methods

### Setting and Participants

This study took place within the context of a larger study examining the effectiveness of the KinderTEK iPad math learning program as a math supplement to core instruction. The parent study was conducted in 13 classrooms, across four elementary schools and three Pacific Northwest districts during the 2017-2018 school year. All kindergarten students (*N* = 343) in participating classrooms were invited to participate. Seven classrooms (189 students) were randomly assigned to treatment and six classrooms (154 students) were randomly assigned to control. Participants in this number line validation sub-study were drawn from the 154 students in the control classrooms (approximaltey half of the sample) ) so that results are not impacted by exposure to the KinderTEK supplemental program. A math battery was administered in the fall (pretest) and spring (posttest) of kindergarten.

Of the 154 students, 27 (18%) had missing testing data on the outcome measure (i.e., spring NSB scores) and were excluded from the current study. Reasons for missing data included students moving out of a participating classroom partway through the year (*n* = 12), students joining a participating classroom partway through the year (*n* = 8), or various other reasons (e.g., absent for testing, language barrier, technical issue; *n* = 7). Welch independent two-sample *t*-tests were conducted to determine if student pretest mathematics scores differed for included students as compared to excluded students with available pretest data. There were no significant differences on any of the pretest mathematics measures: ASPENS, *t*(14.23) = -0.29, *p* = .776; NLA 0-100, *t*(13.72) = -0.73, *p* = .48; and NSB, *t*(15.75) = -0.73, *p* = .48.

Demographic information was missing for 4 (3%) of the 127 students that comprised the analytic sample. Of the 123 students with demographic information: 64 (52%) were female; 103 students (84%) were White, 9 students (7%) were two or more races, 3 students (2%) were Asian, 3 students (2%) were American Indian/Alaskan Native, 1 student was black (0.8%), 1 student was Hispanic (0.8%); 23 students (19%) identified as Hispanic or Latino; 18 students (15%) were English learners; 16 students (13%) received special education services.

## Measures

Three of the parent study's mathematics measures were chosen to investigate the research questions in this sub-study: ASPENS, NSB, and NLA 0-100.

**ASPENS**. The ASPENS assessment is a standardized, individually-administered screening assessment consisting of three one-minute timed subtests. In kindergarten, ASPENS is designed to assess early number sense skills including saying the name of numerals (Number Identification), comparing two numerals and determining which is greater (Magnitude Comparison), and identifying the missing numeral in a string of three numerals (Missing Number). Subtest scores are calculated and weighted to form an overall ASPENS composite score, with greater weight given to subtests with lower raw scores (i.e., Missing Number). The composite score is calculated by combining the raw score from Number Identification, the raw score of Magnitude Comparison x 1.7, and the raw score of Missing Number x 2.7. The authors report test-retest reliability ranging from .71 to .90. Concurrent and predictive validity with the TerraNova Third Edition is reported as ranging from .57 to .63. In the current study, the ASPENS composite score was used to represent a standard kindergarten mathematics screening battery.

**NSB.** The NSB includes 33 items assessing counting knowledge, number identification, magnitude comparison, nonverbal addition and subtraction, and story problems. The measure is untimed and takes approximately 15-20 minutes to administer. Items are scored as correct (1) or incorrect (0) for a total of 33 possible points. The NSB has strong internal consistency (coefficient alpha is reported at .80) and kindergarten performance on the measure was found to be predictive of mathematics achievement at first to third grade (Jordan, Glutting, et al., 2008). In a kindergarten sample, test-retest reliability across students' kindergarten year was found to be adequate (.78 to .81; Jordan, Glutting, & Ramineni, 2010). Diagnostic utility statistics of the

NSB across various time points in kindergarten show that the AUC ranges from .80 to .86, using a mathematics achievement test administered in third grade as the outcome measure (Jordan, Glutting, Ramineni, & Watkins, 2010).

**NLA 0-100.** The NLA 0-100 is a researcher-developed, individually administered measure given on the iPad. It is based off of the number line measure developed by Siegler and colleagues. During the task, the iPad screen displays a horizontal line with endpoints 0 and 100. A female voice explains the concept of the number line and gives the student the practice numerals “0” and “100” to place along the number line, along with affirmative or corrective feedback. Next, the student is asked to drag 26 target numerals appearing one at a time at the top of the screen between 0 and 100 on the number line. Numerals were presented in a random order for each student at pretest and posttest and included 2, 3, 6, 7, 11, 14, 15, 19, 21, 23, 24, 28, 32, 36, 44, 47, 51, 58, 63, 69, 72, 76, 84, 87, 91, and 98.

Upon administration, data collectors and students wore headphones with splitters to verify that students could hear the instructions. The app tracks the accuracy of numeral placement as an absolute error score (for example, if the target numeral was 82 and the student placed it at the actual value of 75.5, the absolute error would be 6.5). The mean absolute error was computed, averaging across all administered trials. Using the mean absolute error, lower scores on the NLA 0-100 (i.e., less overall error) correspond with better performance on the task. The task takes approximately five minutes to administer (Authors, 2018). Due to testing time constraints within schools, data collectors ended the assessment at exactly five minutes, regardless of whether students had finished.

## Procedures

All procedures were approved by the participating districts and the University's IRB. Parent/guardian information letters with opt-out options were delivered two weeks before the study start date and student assent was procured during the initial assessment session. A data collection team were trained to administer the assessments. The training included information about administration logistics (e.g., study procedures, technical troubleshooting for the NLA 0-100), practice with assessments and in-training reliability (i.e., fidelity of administration) for the ASPENS. The initial ASPENS assessment orientation was delivered online with an online checkout. A subsequent in-person training included an in-person reliability checkout. During training, all data collectors had to meet a standard of 90% reliability. For all data collectors, retraining to 100% reliability took place in the days following the training prior to administration of measures to students in schools.

Initial assessments, including the NLA 0-100, NSB, and ASPENS, were administered in October of 2017. The project coordinators (i.e., veteran data collectors) shadow scored assessors' first ASPENS administrations and conducted informal, in-field observations to verify in-field reliability of administration for both the ASPENS and NLA 0-100. Prior to post-testing, a refresher training was provided to all data collectors and new staff were trained and completed reliability checkouts. Posttests included the same measures as pretest and were administered in May of 2018.

## Analyses

To address the first research question (RQ1), Pearson's $r$ bivariate correlations were estimated among the NLA 0-100 and established early numeracy measures. To address the second research question (RQ2), two separate linear regression models were fitted to regress students' spring NSB scores on (a) fall ASPENS scores, or (b) fall NLA 0-100 scores. $R^2$ values for each model were compared to determine which predictor explained more variance in the spring NSB scores. The root mean square error (*RMSE*; the distance between the predicted score and the observed score, squared, and averaged over every observation) was reported for each model to provide context for how well each fall measure functioned as a predictor.

To address the third research question (RQ3), receiver operating characteristic (ROC) analyses were conducted to assess diagnostic accuracy for (a) fall ASPENS scores, and (b) fall NLA 0-100 scores, to determine their classification performance on the spring NSB as a categorical variable defining "risk." To determine "risk", a cut score of 20 was selected on the spring NSB, where a student scoring at or below 20 was considered to be "at risk", and a student scoring above 20 was considered to be "not at risk". We chose a score of 20 as the cut score because (a) it aligned with previous research that suggested that this was the optimal spring Kindergarten NSB cut score (Jordan, Glutting, Ramineni, & Watkins, 2010), and (b) it aligned with the 25th percentile of our sample, which reasonably approximates a percentile score that a school would use in practice to define students who are at risk for poor math outcomes and are potential targets for intervention. We acknowledge that this classification is not absolute, in fact has meaningful associated measurement error, and that any cut score will have these inherent limitations.

ROC analysis uses all possible cut scores of a predictor and visualizes sensitivity (true positive rate, or correct identification of students "at risk") on the *y*-axis and 1 – specificity (false positive rate, or incorrect identification of students "at risk") on the *x*-axis. The metric to evaluate a ROC curve is the AUC (area under the curve), which is on a scale of 0 to 1 where values near 1.0 are best and values near .50 indicate classification is as good as chance.

The AUCs of the two ROC models (fall ASPENS and fall NLA 0-100) were compared with a bootstrap test for two correlated ROC curves where 2,000 bootstrap samples are drawn from the data, each containing exactly the same number of "at risk" and "not at risk" students as the original sample. For each bootstrap sample, the AUC of the two ROC curves and their difference are computed, and the following formula was used to compute a $D$ statistic:

$$D = \frac{AUC_1 - AUC_2}{s}$$

where $s$ is the standard deviation of the bootstrap differences and $AUC_1$ and $AUC_2$ are the AUCs of the original fall ASPENS and fall NLA 0-100 ROC curves.

To accompany the ROC analyses, two separate logistic regression models were conducted, regressing students' spring NSB risk status on (a) fall ASPENS scores, and (b) fall NLA 0-100 scores, to generate a predicted probability of "risk" for each student, based on the model. These predicted probabilities are used to make class predictions, where a threshold is established (anywhere between 0 and 1, often at 0.5) to classify everything above the threshold as "at risk" and everything below as "not at risk."

All analyses were conducted in R (R Core Team, 2019), with the following packages: ggthemes (Arnold, 2019); here (Müller, 2017); Hmisc (Harrell et al., 2019); knitr (Xie, 2020); janitor (Firke, 2019); modelr (Wickham, 2019); pROC (Robin et al., 2011); rio (Chan, Chan, Leeper, & Becker, 2018); and tidyverse (Wickham et al., 2019).

## Results

**Distributions of Mathematics Measures.** The spring NSB scores were not normally distributed and had a negative skew. The scores were squared to more approximate a normal curve, but the transformation did not meaningfully change the results, so the original scale was used. The fall ASPENS scores were also not normally distributed and were positively skewed.

We took the square root of the scores and found that the resulting distribution better approximated a normal distribution. This transformation did affect the results of the linear models, improving the relation with the spring NSB scores. Therefore, the square root of the scores was used in the analyses for RQ1 and RQ2.

**Addressing the NLA 0-100 Five Minute Cut-off.** With the five minute cut-off on the NLA 0-100, on average students completed 20.37 (*SD* = 5.44) of the 26 items. Of the 127 students comprising the final sample, 39 students (31%) completed all 26 items on the fall NLA 0-100. All students completed at least 8 of the 26 items. The correlation between the number of items completed on the fall NLA 0-100 and spring NSB scores was examined and found to be low (*r* = .20). Additional follow-up sensitivity analyses were conducted to determine whether the number of items completed on the NLA 0-100 influenced the results. That is, all analyses were repeated for fall NLA 0-100 scores using the number of items completed as a covariate, as well as using a composite fall NLA 0-100 (fall NLA 0-100 / number of items completed). Including the number of items completed did not result in a better model of the spring NSB data, and thus were excluded from the results presented here. The results of the linear regression model including the number of items completed as a covariate are shown as a reference in Table 2.

**RQ1.** Descriptive statistics and correlations are displayed in Table 1. From fall to spring of kindergarten, student performance improved across all measures. The decrease in NLA 0-100 scores from fall to spring corresponds to less error in numeral placement. Correlations among all measures from fall to spring were moderate and all significant at the *p* < .01 level. All NLA 0-100 scores were negatively correlated with the other early numeracy measures given that lower scores on the NLA 0-100 corresponded to less error and thus reflected better performance. The predictive validity correlation between the fall NLA 0-100 and the spring NSB was -.58, and the predictive validity correlation between the fall ASPENS and the spring NSB was slightly higher at .63.

**RQ2.** The results of the linear regression models can be seen in Table 2. Because there was only one predictor in each model, the $R^2$ was simply the squared correlation showed in Table 1. The regression model with the square root of the fall ASPENS was the best model, as it has the lowest *RMSE* (3.92) and the highest $R^2$ (.51). The untransformed fall ASPENS scores were still a better predictor of spring NSB scores than the fall NLA 0-100, which had an *RMSE* of 4.68 and an $R^2$ of .34. The inclusion of the number of items completed on the fall NLA 0-100 only slightly increased the model prediction. Note that the *SD* of the spring NSB scores was 5.77, so that the mean error around the predictions of the best model, the square root of the fall ASPENS, was about 0.68 of a *SD*, a fairly large error.

**RQ3.** Figure 1 shows the ROC curves for both the fall ASPENS and the fall NLA 0-100, as well as their respective AUCs. (Note that the untransformed fall ASPENS scores were used for this analysis given that the square root transformation does not affect the results of the ROC.) The AUC of the ASPENS was .95, with a 95% CI from .90 to .98. The AUC of the fall NLA 0-100 was .78, with a 95% CI from .68 to .86. The bootstrap test for two correlated ROC curves yielded a *D* statistic of 3.26, *p* < .001, indicating that the ASPENS AUC was significantly higher than the NLA 0-100 AUC, and therefore had better “predictive accuracy” than the fall NLA 0-100.

Based on the ROC analyses, two decision rules were used to examine sensitivity and sensitivity and are shown in Table 3. First, sensitivity was set at or near .90 (Johnson et al., 2009; Seethaler & Fuchs, 2010) and associated specificities and cut scores on the fall ASPENS and NLA 0-100 were examined. In school practice, correctly classifying students that are at risk is often prioritized, and thus this decision rule was used to place the greatest emphasis on accurate classification of students at-risk. Using this approach, with a sensitivity of .91, the fall ASPENS had a specificity of .83 (cut score = 22.50), whereas the fall NLA 0-100 had a specificity of .39 (cut score = 26.22). Thus, while correctly identifying 91% of students “at risk”, the fall ASPENS correctly identified 83% of students “not at risk”, whereas the fall NLA 0-100 only correctly identified 39% of students “not at risk”.

Second, both sensitivity and specificity were maximized by identifying the highest sum of the two values (closest to 2.0; Author et al., 2011b). The fall ASPENS had a sensitivity of .91 and a specificity of .83 (cut score = 22.50), whereas the fall NLA 0-100 had a sensitivity of .69 and a specificity of .80 (cut score = 37.44).

Figure 2 shows the density (distribution) of predicted spring NSB class probabilities (“at risk” or “not at risk”) from the logistic regressions with either the fall ASPENS or the fall NLA 0-100 as a predictor. The horizontal line at .50 represents a common threshold for bifurcating the probabilities into class membership, but any threshold can be applied in practice. If the “at risk” and “not at risk” densities were completely separated, the threshold would serve as a point between the densities and classification would be perfect. The more the densities overlap, the greater the misclassification. Figure 2 shows a greater separation of densities for the fall ASPENS, which represents the high AUC shown in the ROC analysis, and good sensitivity and specificity. For the fall NLA 0-100, the “at risk” distribution is relatively uniform across the predicted probabilities, meaning there is no point on the x-axis where the two densities are most clearly separated, resulting in a lower AUC and poor specificity when sensitivity is set at/near .90.

## Discussion

The mental number line has been proposed as a central conceptual structure underlying the development and application of early numeracy skills. Studies in the developmental and cognitive literature have shown strong relationships between performance on number line assessments and student mathematics achievement, but to date, only one study has explored the number line assessment as an early numeracy screening measure. The current study adds to this emerging literature base by investigating the number line measure using linear regression and classification accuracy analyses and contextualizes results as compared to a set of established early numeracy CBMs.

Similar predictive validities were found between the fall ASPENS ($r$ = .63) and fall NLA 0-100 ($r$ = -.58) scores with students' spring NSB scores (RQ1). Correlations of the NLA 0-100 measure in general were significant and moderate, though marginally lower than correlations of the ASPENS with other measures at various time points. Linear regression analyses (RQ2) revealed that the transformed fall ASPENS scores resulted in the best overall model, with a *RMSE* of 3.92 and the highest $R^2$ value of .51 (i.e., the ASPENS individually explained 51% of the variance in students' spring NSB scores). In comparison, the fall NLA 0-100 had a higher *RMSE* of 4.68 and a lower $R^2$ of .34 (i.e., explaining 34% of the variance in students' spring NSB scores). Thus, the fall NLA 0-100 measure fared worse as a predictor of mathematics outcomes in the spring of kindergarten.

Using ROC analyses (RQ3), the ASPENS was overall a better classifier of students at-risk and not at-risk, using a cut score of 20 (25th percentile of the current sample) on the spring NSB to designate risk status. The AUC of the ASPENS was significantly greater at .95 as compared to the AUC of the NLA 0-100 at .78. The AUC can be interpreted as the probability of distinguishing between one randomly selected student from the at-risk sample, and one randomly selected student from the not at-risk sample (Rodrigues, Jordan, & Hansen, 2019). Therefore, the ASPENS would correctly identify student risk status 95% of the time, whereas the NLA 0-100 would correctly identify risk status 78% of the time. While the ASPENS significantly outperformed the NLA 0-100 in accurately predicting risk, it should still be noted that both measures meet the recommendation for screening measures in educational research, to meet a minimum threshold of an AUC of .75 or above (Cummings & Smolkowski, 2015).

Both the ASPENS and the NLA 0-100 were further examined in two ways to compare and contrast the utility of an established screening measure (i.e., the ASPENS) compared to one that is relatively unexplored in the mathematics screening literature. First, sensitivity was set at or near .90 based on recommended guidelines for evaluating screening measures (Johnson et al., 2009; Seethaler & Fuchs, 2010). Using this approach, the ASPENS outperformed the NLA 0-100, with a specificity of .83 compared to a specificity of .39. A second approach is to select a cut score that maximizes both sensitivity and specificity (Authors et al., 2014). With this approach, the ASPENS maintained a sensitivity of .91 and specificity of .83, whereas the NLA 0-100 resulted in a decrease in sensitivity to .69, and an increase in specificity to .81.

When interpreting these findings, it is important to acknowledge the limitations of the current study. First and foremost, we had a relatively small sample size across three districts in the Pacific Northwest. The demographics of our sample reflected the larger demographics of this region, which included a majority of students that were White, with some diversity in terms of students that were Hispanic or Latina/o and English learners. Future research should be conducted in other regions of the country with a more diverse sample. Second, due to the longitudinal nature of our study, approximately 18% of our sample had missing data from the primary outcome of interest, and were excluded from the analyses. In considering this limitation, it should be noted that no significant differences were found in mathematics pretest scores on any study measures between students who left the study and those with data at both time points. Third, a five-minute cut-off procedure was applied for the NLA 0-100 in efforts to keep assessment time reasonable. As a result, on average students completed about 20 of the 26 items. Sensitivity analyses were conducted and revealed that including the number of items completed (out of 26) as a covariate in the regression models did not result in a better model of the spring NSB data. Last, in conducting ROC analyses, a continuous outcome measure (i.e., students' spring NSB scores) was dichotomized by selecting a cut score aligned with recommendations from research and the 25th percentile of our sample to designate risk status. One challenge with using classification accuracy approaches in educational research is that there is no "true" result for risk in an academic area. When using classification accuracy analyses, academic risk must be defined on a criterion measure, and various performance levels representing risk have been utilized by different researchers, including the 16th, 25th, and 40th percentiles (Author et al., 2011b; Seethaler & Fuchs, 2010). When selecting a cut score to designate risk on an outcome measure and interpreting the results of ROC analyses, schools should consider how risk was defined on the criterion along with considering issues specific to the study sample including the base rate of risk in the sample population.

The extent to which it would be worthwhile to include a measure such as the NLA 0-100 in screening batteries is up for debate and the current study's findings must be considered in light of present practice and implications for schools. At a minimum, the value of new screeners should be weighed from a cost-benefit perspective such that the psychometric properties of number line tasks are viewed in light of their ability to be used efficiently and effectively in instructional decision making including both screening and progress monitoring decisions. Given the results of the present study, we do not consider modifications to current screening approaches to be justified. However, given the hypothesized role of the number line as a central construct in early mathematics development and the findings of the current study, the number line warrants further exploration and additional research in targeted areas.

While the AUC of the NLA 0-100 was above the recommended threshold of .75 (Cummings & Smolkowski, 2015), it fared considerably worse than the ASPENS when examining sensitivity and specificity statistics. Given that the NLA 0-100 overidentified students students later found to be not at-risk, it is possible that a different type of number line task may be better suited for kindergarten students. Some researchers theorize that the number line measure may be tapping into more advanced mathematical skills, such as proportional reasoning and division (Cohen & Sarnecka, 2014). For example, when placing the numerals along a number line spanning 0 to 100, more successful students are theorized to use benchmark numerals such as 25, 50, and 75 to divide the number line into quarters and make their estimate. Students in kindergarten, especially those at-risk, may not have developed this skill. Cohen and Sarnecka (2014) administered a number line assessment that did not have an upper bound, but did have a line segment showing the distance between 0 and 1. The structure of this task required students to iterate the length of the segment to determine placement of target numbers (e.g., to place the numeral “13”, iterating the line segment from 0 to 1, 13 times). Because these skills are more aligned with how kindergarten students may approach the number line task, similar approaches relying upon line segments that students can iterate, or structured number lists with missing values, may be better suited for screening purposes. At the very least, number line tasks constructed in varying ways warrant exploration as part of early numeracy screening batteries.

One unique and potentially promising feature of the number line estimation task as a screening tool is its relative robustness across grades and mathematical content. In the upper elementary grades, fraction number lines (e.g., students place fractions on a number line with endpoints from 0 to 2) in particular were found to have strong relations with mathematical performance (Schneider et al., 2018). This suggests that students may be relying upon similar mental number line representations to build their understanding of fraction magnitudes. Given the findings across whole and rational number systems (Schneider et al., 2018), and drawing upon the broader idea that students use a mental number line to make sense of number systems overall (Siegler et al., 2011), the number line task may lend itself to serve as a common screener across grade levels.

Additionally, while typical math screeners are bound to certain ages or content, supplementing screening batteries with the number line assessment may allow for advantageous comparisons of performance across years on relatively similar tasks. For example, Rodrigues et al. (2019) found that combining fraction concepts items and fraction number line items led to better prediction models in the 4th, 5th, and 6th grades compared to either set of items alone. The researchers also used best subset automatic linear modeling to reduce the total pool of fraction concept and fraction number line items to only include combinations of items with the best classification accuracy. Future number line research should investigate these types of novel approaches to allow for more accurate and efficient screening.

## Conclusion

The research on screening using number line measures is still in its infancy and necessitates further investigation. By more directly assessing the conceptual structures that students use to make sense of number systems, the field may develop screening tools that tap into underlying mathematical skills and meet calls to develop and investigate new early numeracy screening assessments. As students grow in their mathematical proficiency and develop understanding of new number systems (e.g., whole numbers and the base ten system in the early elementary grades, rational numbers in the late elementary grades), different types of number line measures aligned with student grade and skill level may enable differentiation between students at risk and on track in mathematics. For kindergarten students making the transition from informal to formal number knowledge, it is of the utmost importance that struggling students are identified early on to enable extra supports prior to students falling further behind. Given the disparities in early exposure to mathematics across socioeconomic lines and the resulting impact on mathematics achievement at kindergarten entry, the need for accurate screening measures is all the more salient. Developments in this line of research may lead to deeper understanding of how students progress in their mathematical thinking, along with increasing our knowledge of how to build comprehensive multi-tiered systems that can efficiently and effectively promote mathematical learning for all students.

# References

Alanzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system.* Eugene, OR: Center for educational Assessment Accountability. Retried from http://easycbm.com.

Albers, C. A., & Mission, P. L. (2014). Universal screening of english language learners: Language proficiency and literacy *Universal screening in educational settings: Evidence-based decision making for schools.* (pp. 275–304). Washington, DC, US: American Psychological Association.

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*(2), 231–244. doi: https://doi.org/10.1016/j.ecresq.2011.08.003

Arnold, J. B. (2019). *ggthemes: extra themes, scales and geoms for 'ggplot2'.* R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes

Author et al. (2004).

Author et al. (2008).

Author et al. (2011a).

Author et al. (2011b).

Author et al. (2014).

Authors. (2018).

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

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38*, 333–339. doi: 10.1177/00222194050380040901

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

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

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development, 61*, v–265.

Chan, C., Chan, G., Leeper, T. J., & Becker, J. (2018). *rio: A Swiss-army knife for data file I/O.* R package version 0.5.16.

Chard, D. J., Clarke, B., Baker, S. K., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*, 3–14. doi: 10.1177/073724770503000202

Clements, D. H., Sarama, J., & DiBiase, A. M. (2003). *Engaging young children in mathematics. Standards for early childhood mathematics education*. New York: Routledge.  doi: 10.4324/9781410609236.

Cohen, D. J., & Sarnecka, B. W. (2014). Children’s number-line estimation shows development of measurement skills (not number representations). *Developmental psychology, 50*(6), 1640–1652. doi: 10.1037/a0035901

Conoyer, S. J., Foegen, A., & Lembke, E. S. (2016). Early Numeracy Indicators: Examining Predictive Utility Across Years and States. *Remedial and Special Education*, *37*(3), 159-171. doi: 10.1177/0741932515619758

Cummings, K. D., & Smolkowski, K. (2015). Bridging the gap: Selecting students at risk of academic difficulties. *Assessment for Effective Intervention*, 41, 55–61. doi:10.1177/1534508415590396

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232. doi: 10.1177/001440298505200303

Firke, S. (2019). *janitor: Simple tools for examining and cleaning dirty data.* R package version 1.2.0. https://CRAN.R-project.org/package=janitor

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computation skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology,* 100, 30–47.

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development*, *78*(4), 1343-1359.

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*, 277–299. doi: 10.1080/87565640801982361

Gersten, R. M., & Chard, D. J. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *Journal of Special Education, 33,* 18–28. doi: 10.1177/002246699903300102

Gersten, R. M., Clarke, B., Jordan, N., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78*, 423–445. Retrieved from http://cec.metapress.com/content/B75U2072576416T7

Gersten, R. M., Clarke, B., & Mazzocco, M. M. M. (2007). Historical and contemporary perspectives on mathematical disabilities. In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 7–29). Baltimore, MD: Paul H. Brookes Pub. Co.

Gersten, R. M., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38,* 293–304. doi: 10.1177/00222194050380040301

Ginsburg, H., & Baroody, A. J. (2003). *TEMA-3: Test of early mathematics ability*. Pro-ed.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education, 22*(3), 170–218. doi: 10.2307/749074

Griffin, S. (2004). Building number sense with number worlds: A mathematics program for young children. *Early Childhood Research Quarterly, 19*(1), 173–180. doi: 10.1016/j.ecresq.2004.01.012

Griffin, S., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 24–49). Cambridge, MA: MIT Press.

Hampton, D. D., Lembke, E. S., Lee, Y. S., Pappas, S., Chiong, C., & Ginsburg, H. P. (2012). Technical adequacy of early numeracy curriculum-based progress monitoring measures for kindergarten and first-grade students. *Assessment for Effective Intervention*, *37*(2), 118-126. doi: 10.1177/1534508411414151

Harrell, F. E. Jr, Dupont, C. et al. (2019). *Hmisc: Harrell miscellaneous*. R package version 4.3-0. https://CRAN.R-project.org/package=Hmisc

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments?. *Learning Disabilities Research & Practice*, *24*(4), 174-185. doi:10.1111/j.1540-5826.2009.00291.x

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Ed.), *Mathematical difficulties: Psychology and intervention* (pp. 45–57). San Diego, CA: Academic Press.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and individual differences, 20*, 82–88. doi: 10.1016/j.lindif.2009.07.004

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, *39*(2), 181-195. doi: 10.1080/02796015.2010.12087772

Jordan, N. C., Hansen, N., Fuchs, L. S., Siegler, R. S., Gersten, R., & Micklos, D. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology, 116*(1), 45–58. doi: https://doi.org/10.1016/j.jecp.2013.02.001

Judge, S., & Watson, S. M. R. (2011). Longitudinal outcomes for mathematics achievement for students with learning disabilities. *The Journal of Educational Research, 104*, 147–157. doi: 10.1080/00220671003636729

Kalchman, M., Moss, J., & Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

LeFevre, J. A., Skwarchuk, S.-L., Smith-Chant, B. L., Fast, L., Kamawar, D., & Bisanz, J. (2009). Home numeracy experiences and children's math performance in the early school years. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 41*(2), 55–66. doi: 10.1037/a0014532

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice, 24*(1), 12–20. doi: 10.1111/j.1540-5826.2008.01273.x

Melhuish, E. C., Phan, M. B., Sylva, K., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2008). Effects of the home learning environment and preschool center experience upon literacy and numeracy development in early primary school. *Journal of Social Issues*, *64*(1), 95–114. doi: 10.1111/j.1540-4560.2008.00550.x

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306–321. doi: 10.1177/0022219408331037

Müller, K. (2017). *here: A simpler way to find your files.* R package version 0.1. https://CRAN.R-project.org/package=here

Okamoto, Y., & Case, R. (1996). II. Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for research in Child Development*, *61*(1-2), 27-58.

Pedhazur, E. J. (1998). *Multiple regression in behavioral research: Explanation and prediction*. Belmont, CA: Wadsworth,.

R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12, p. 77.* doi: 10.1186/1471-2105-12-77

Rodrigues, J., Jordan, N. C., & Hansen, N. (2019). Identifying Fraction Measures as Screeners of Mathematics Risk Status. *Journal of learning disabilities, 52*(6), 480-497. doi: 10.1177/0022219419879684

Saxe, G. B., Guberman, S. R., & Gearhart, M. (1987). Social processes in early number development. *Monographs of the Society for Research in Child Development, 52*(2), 162–162. doi: 10.2307/1166071

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. doi: 10.1111/cdev.13068

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children*, *77*(1), 37-59.

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., & Opfer, J. E. (2003). The development of numerical estimation. *Psychological Science, 14*, 237–250. doi: doi:10.1111/1467-9280.02438

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*, 273–296. doi: 10.1016/j.cogpsych.2011.03.001

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*, 99–120. doi: 10.1016/j.ecresq.2004.01.002

The Psychological Corporation. (2002a). Early math diagnostic assessment. San Antonio, TX: Author.

Wickham, H. (2019). *modelr: Modelling functions that work with the pipe.* R package version 0.1.5. https://CRAN.R-project.org/package=modelr

Witzel B. S., Riccomini, P. J., & Herlong, M. L. (2013). *Building number sense through the Common Core.* Thousand Oaks, CA: Corwin.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III tests of achievement.

Xie, Y. (2020). *knitr: A general-purpose package for dynamic report generation in R.* R package version 1.27.

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

Table 1

*Descriptive Statistics and Correlations Among Measures (Fall 2017 and Spring 2018)*

| | 1 | 2 | 3 | 4 | 5 | Mean (*SD*) |
|---|---|---|---|---|---|---|
| 1. Fall NSB | -- | | | | | 16.07 (6.70) |
| 2. Fall ASPENS | .83 | -- | | | | 45.12 (40.44) |
| 3. Fall NLA 0-100 | -.57 | -.62 | -- | | | 33.03 (11.39) |
| 4. Spring NSB | .70 | .63 | -.58 | -- | | 24.10 (5.77) |
| 5. Spring ASPENS | .61 | .71 | -.54 | .63 | -- | 100.45 (48.10) |
| 6. Spring NLA 0-100 | -.52 | -.54 | .58 | -.56 | -.56 | 24.28 (11.78) |

*Note.* All values significant at *p* < .01. *SD* = standard deviation.

Table 2

*RMSE and Explained Variance ($R^2$) of the Spring NSB from Linear Regression Models*

| Predictors | *RMSE* | $R^2$ |
|---|---|---|
| sqrt(Fall ASPENS) | 3.92 | .51 |
| Fall ASPENS | 4.33 | .40 |
| Fall NLA 0-100 + *n* item completed | 4.62 | .35 |
| Fall NLA 0-100 | 4.68 | .34 |

Table 3

*Classification Accuracy for Fall Screenings with Sensitivity at/near .90 and Maximizing both Sensitivity and Specificity*

| Fall ASPENS | | | Fall NLA 0-100 | | |
|---|---|---|---|---|---|
| Cut Score | Sensitivity | Specificity | Cut Score | Sensitivity | Specificity |
| 22.50 | .91 | .83 | 26.22 | .91 | .39 |
| 22.50 | .91 | .83 | 37.44 | .69 | .81 |

Figure 1

*ROC Curves for Fall ASPENS and Fall NLA 0-100*

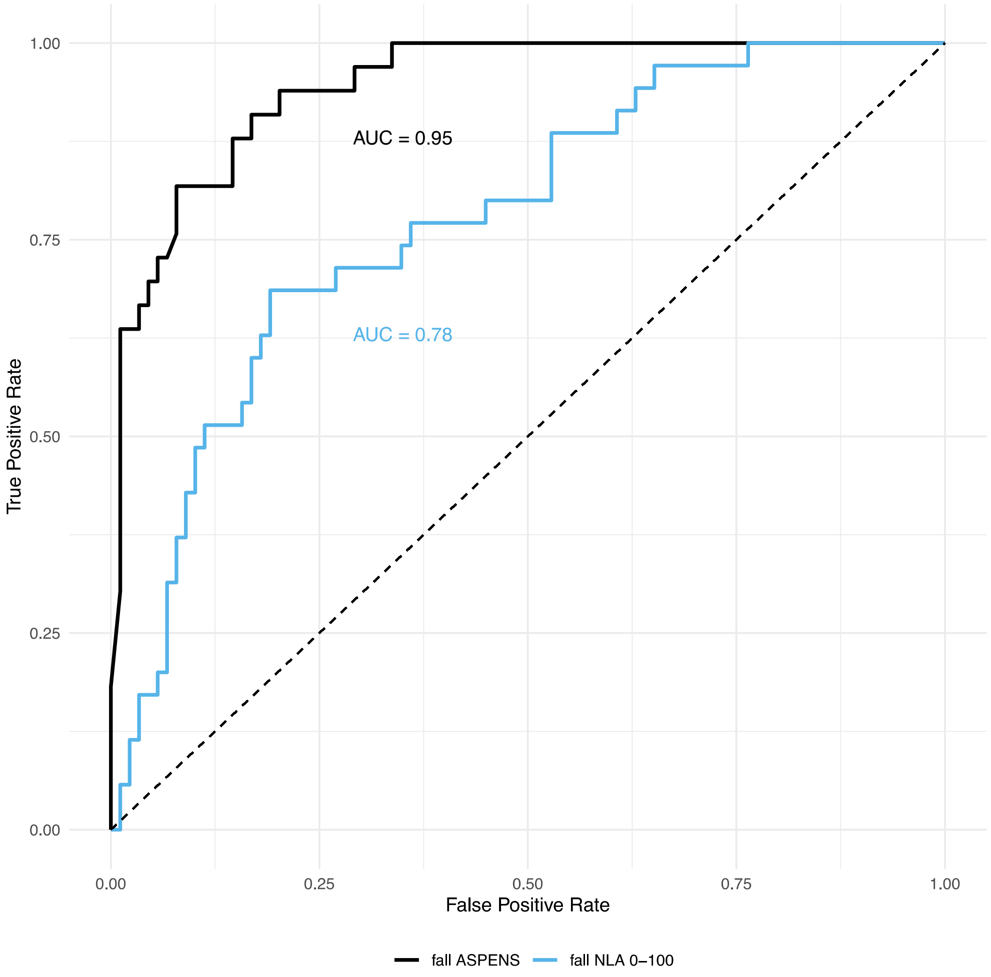
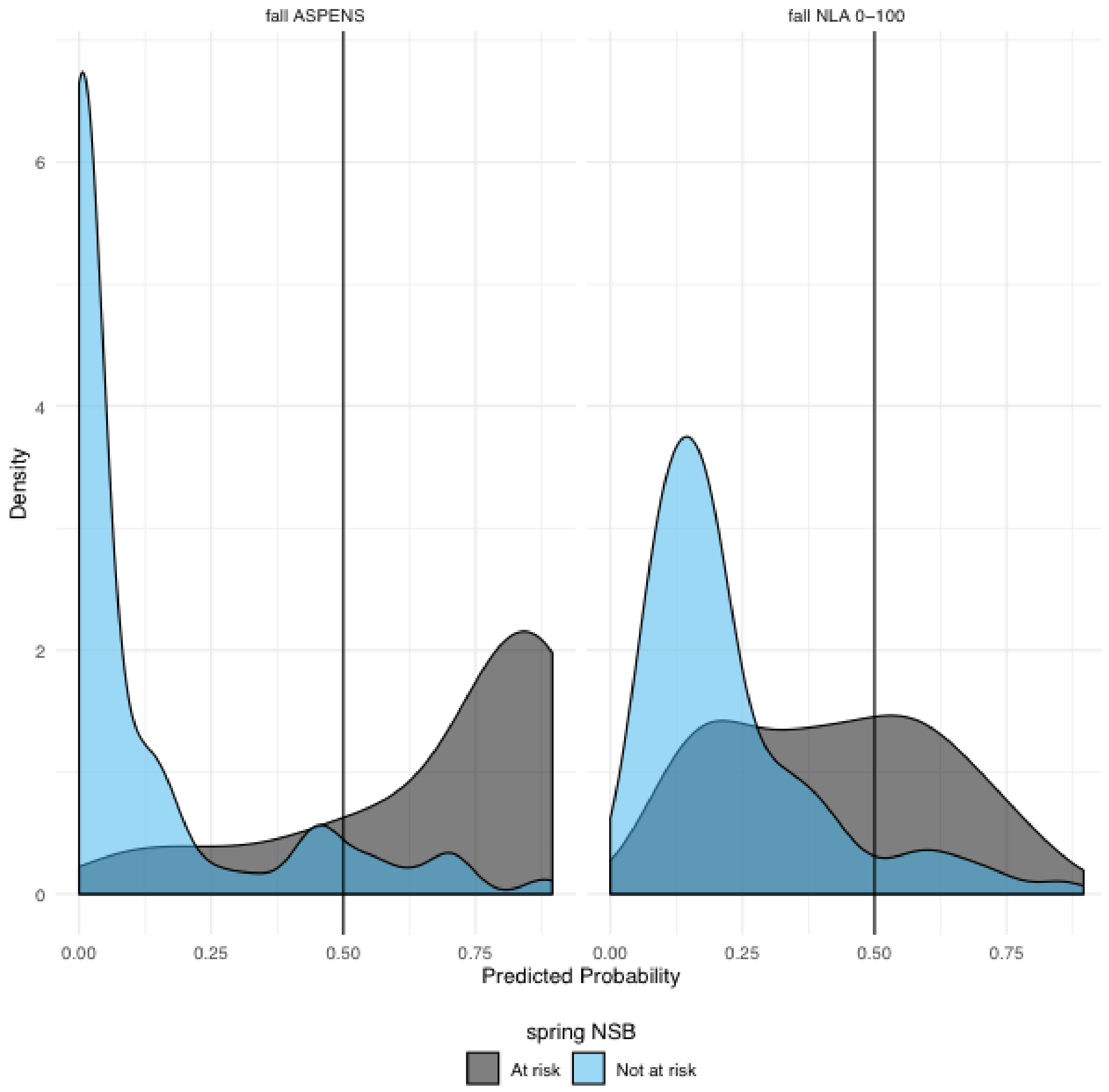


Figure 2

*Densities of Predicted Spring NSB Class Probabilities*


fall ASPENS
fall NLA 0–100
Density
Predicted Probability
spring NSB
At risk
Not at risk