Exploring an early numeracy screening measure for English learners in primary grades

Tasia Brafford
The Meadows Center for Preventing Educational Risk at The University of Texas at Austin

Ben Clarke
University of Oregon
Center on Teaching and Learning

Russell M. Gersten
Instructional Research Group

Keith Smolkowski
Oregon Research Institute

Marah Sutherland
University of Oregon
Center on Teaching and Learning

Joe Dimino
Instructional Research Group

David Fainstein
University of Oregon
Center on Teaching and Learning

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324B200012 to The University of Texas at Austin. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Citation:

Brafford, T., Clarke, B., Gersten, R. M., Smolkowski, K., Sutherland, M., Dimino, J., & Fainsten, D. (2023). Exploring an early numeracy screening measure for English learners in primary grades. *Early Childhood Research Quarterly, 63*, 278-287. https://doi.org/10.1016/j.ecresq.2022.12.007

Early Childhood Research Quarterly 63 (2023) 278–287

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

journal homepage: www.elsevier.com/locate/ecresq

Exploring an early numeracy screening measure for English learners in primary grades

Tasia Brafford[a,*], Ben Clarke[b], Russell M. Gersten[c], Keith Smolkowski[d], Marah Sutherland[b], Joe Dimino[c], David Fainstein[b]

[a] *University of Texas at Austin/Meadows Center for Preventing Educational Risk, 1912 Speedway, Stop D4900, SZB 5.110, Austin, TX 78712*
[b] *University of Oregon/Center on Teaching and Learning, 1600 Millrace Dr. Suite 207, Eugene, OR 97403*
[c] *Instructional Research Group, 4281 Katella Ave # 205, Los Alamitos, CA 90720*
[d] *Oregon Research Institute, 1776 Millrace Dr, Eugene, OR 97403*

ARTICLE INFO

Keywords:
English learners
Screening
At risk
Mathematics
Early mathematics

ABSTRACT

We investigated the technical characteristics of a brief early numeracy screening battery for both English learners (ELs) and English proficient students (EPs). Results indicated there were differences in performance of ELs and EPs. Further, we found reasonable overall accuracy of the screener predicting student outcomes. Similar overall accuracy results were found for ELs and EPs, as well as for predicting academic performance. We discuss study results related to sensitivity, specificity, and negative and positive predictive power as they relate to implications for practice, including screening for risk for ELs and the challenge of false positives in screening systems. We conclude by proposing future avenues of research.

In the U.S., 4.8 million school-aged students (10%) are English learners (ELs; U.S. Department of Education, 2017; U.S. Department of Education, National Center for Education Statistics, 2017). As the fastest growing student population in U.S. schools, the population of ELs is expected to continually grow, with 40% of school-aged children projected to be ELs by 2030 (Calderón et al., 2011; Thomas & Collier, 2002). In most states, the percentage of ELs increased between the 2009-2010 to 2014-2015 school years, with five states reporting increases of over 40% (U.S. Department of Education, Office of English Language Acquisition, National Clearinghouse for English Language Acquisition 2017). It is concerning that this population has lower rates of academic achievement than their native English-speaking peers. In 2019, the National Assessment of Educational Progress (NAEP) reported only 16% of Grade 4 students identified as ELs were proficient in mathematics, compared to 44% of English proficient speakers (EPs). Outcomes were worse for Grade 8 ELs, with only 5% scoring at the proficient level, compared to 36% of EPs (NAEP, 2019). Differences detected in Grade 4 begin much earlier, with longitudinal trajectories indicating early mathematics difficulties (i.e., those apparent in preschool and kindergarten) persist throughout elementary and middle school (Morgan et al., 2014).

A focus on early mathematics is critical, given the disheartening longitudinal evidence that documents a pattern in which students with low kindergarten mathematics achievement experience persistent, increasing difficulty in mathematics throughout elementary and middle school (Morgan et al., 2014). The relationship between early and later mathematics has been found to be significantly stronger than relationships with other kindergarten predictors, including cognitive variables, and later mathematics (Morgan et al., 2014). Further evidence demonstrates mathematics achievement difficulties show an intractable pattern of performance (e.g. Bodovski & Farkas, 2007; Duncan et al., 2007; Hanich et al., 2001). While the research on ELs' longitudinal mathematics development is limited, similar patterns have been documented (Reardon & Gallindo, 2009). Support for ELs' mathematics development is complicated by the limited time spent on mathematics instruction in early grades, particularly in kindergarten (La Paro et al., 2009), the corresponding lack of institutional support in mathematics (Clarke et al., 2014), and, specifically for ELs, the additional need to focus instruction on language development.

Despite these challenges, significant advancements have been made in mathematics for the general population of learners, including in key areas such as the development and validation of screeners to identify risk (Fuchs et al., 2009; Gersten et al., 2012; Jordan et al., 2010) and promising interventions in the early elementary grades (Nelson & McMaster, 2019). However, the same strides have not been made with ELs. A meta-analysis of interventions with ELs from 2000 to 2012 indicated that there were no published studies utilizing randomized con-

[*] Corresponding author.
E-mail addresses: brafford@utexas.edu (T. Brafford), clarkeb@uoregon.edu (B. Clarke), rgersten@uoregon.edu (R.M. Gersten), keiths@ori.org (K. Smolkowski), marahs@uoregon.edu (M. Sutherland), joe.dimino@inresg.org (J. Dimino), dfainste@uoregon.edu (D. Fainstein).

https://doi.org/10.1016/j.ecresq.2022.12.007
Received 14 March 2022; Received in revised form 2 November 2022; Accepted 12 December 2022

trol trials in mathematics (Richards-Tutor et al., 2016). Since then, a relatively small number of studies have focused on ELs and mathematics (e.g., Doabler et al., 2016; Doabler et al., 2019; Driver & Powell, 2017). Our review of the literature found no studies investigating kindergarten or Grade 1 mathematics screeners with ELs, though there have been calls for these explorations (Purpura et al., 2015). In literacy contexts, Cummings et al. (2021) explored the use of different thresholds and if they differed for ELs and EPs in kindergarten through Grade 3. Cummings and colleagues (2021) found that similar cut scores across groups were mostly consistent, with some differences in the kindergarten year. Within a multi-tiered systems of support model, the first step in the process of providing adequate and appropriate supports to students in need of mathematics intervention is identification of risk, therefore, research on the utility of screeners with ELs is desperately needed.

1. Early mathematics screening measures

Given longitudinal trajectories of mathematics development for the general population of learners (Jordan et al., 2007), investigations to detect risk early were warranted and have been widely researched (e.g., Fuchs et al., 2007; Gersten et al., 2012). These works focused on how and if students in early grades (i.e., kindergarten and Grade 1) were best identified as at-risk using different types of assessment activities; ELs were part of the population of learners within these investigations, but any differences between ELs and EPs were not specifically highlighted. Brief measures composed of discrete tasks have often been used as screeners, with number sense typically being the focus in the screeners for early grades (Gersten et al., 2012; Mazzocco, 2005). Numerous investigations have established these measures to have predictive validity and classification accuracy, demonstrating that tasks such as number identification, quantity discrimination, and strategic counting have the potential to accurately identify at-risk kindergarten students (Chard et al., 2005; Hampton et al., 2012; Seethaler & Fuchs, 2010).

Measures typically used in early mathematics screening consist of two underlying frameworks: (a) basic readiness skills and (b) number understanding and underlying mathematics concepts. Measures focused on basic readiness skills (e.g., Clarke & Shinn, 2004; Lembke et al., 2008) require students to complete tasks such as rote oral counting or numeral identification. Other screening measures focus on number understanding, such as magnitude comparison and strategic counting, and are theorized to reflect the development of a mental number line and tap into students' general understanding of number (Berch, 2005; Booth & Siegler, 2008; Case & Okamoto, 1996; Fuchs et al., 2007; Gersten & Chard, 1999; Gersten et al., 2012). This central construct of the number line continues throughout elementary school, with evidence that number line performance serves as a strong predictor of algebra readiness and success (Bailey et al., 2014). Further, measures that examine broader performance in the areas of number sense and underlying number knowledge and concepts also require competence with number line understanding and related skills (e.g., Baker et al., 2002; Jordan et al., 2010). Although both frameworks or approaches to early mathematics skills (i.e., basic readiness skills and number understanding and underlying mathematics concepts) have demonstrated strong utility as screening measures, ELs have not been the target population of these investigations. Therefore, measures using these approaches represent logical starting points for investigating screeners with ELs.

One important element of tests in U.S. schools is the conflation of the targeted construct being assessed and skills in the English language. Often, discussions of measurement development and validity do not include the language of administration and simply assume that students speak English with sufficient proficiency to complete the assessment tasks. Commonly used mathematics screening measures assess mathematics knowledge and skills *in English*. Ideally, instruction in early mathematics would explicitly teach the necessary vocabulary in English to increase access to content among ELs. This interaction between the language of instruction and the language of test administration is critical to the understanding of the constructs and the results of assessment for any language-dependent measure.

1.1. Investigation of a universal screener for ELs

In the current study, our objective was to examine psychometric characteristics, predictive validities, and classification accuracy of the *Assessing Student Proficiency of Early Number Sense* (ASPENS; Clarke et al., 2011) screening measure for a sample of EL and EP students in kindergarten and Grade 1. ASPENS measures use the two different frameworks of early mathematics, including activities based on both basic readiness skills (i.e., Number Identification, Basic Arithmetic Facts and Base 10) and number understanding and underlying mathematics (i.e., Magnitude Comparison, Missing Number). ASPENS consists of three timed curriculum-based measures designed to assess students' early number sense. All measures are administered, and students answer as needed, in English. In kindergarten, students identify randomly sampled numbers ranging from 0 to 20 as quickly as possible (Number Identification), name the greater of two written numerals (Magnitude Comparison), and name the missing number in a string of visually presented numbers (Missing Number). In Grade 1, the measures include Magnitude Comparison, Missing Number, and a measure where students write the answer to facts that cross 10 (e.g., 8 + 5; Basic Arithmetic Facts and Base 10). While students are required to verbally state number names in English for three of ASPENS subtests, we argue that the linguistic complexity of these tasks is limited due to the straightforward, scripted assessor directions and the multiple opportunities for students to complete practice items and receive feedback prior to the assessment. Though these tasks are like those typically included in early numeracy screening measures, the extent to which these measures accurately identify ELs compared to EPs is unknown and pertinent for exploration.

1.2. Purpose and research questions

The primary purpose of this research project was to investigate the psychometric properties of ASPENS for ELs in kindergarten and Grade 1. This investigation provides key information on whether screening measures that are commonly used across both EL and EP populations are appropriate for use with ELs. Through this investigation, the field of early childhood education will gain insights into the common practice of using the same screener across all students and initial information on if this practice is appropriate for EL students. Further, this research adds to the necessary literature on the psychometric properties of early mathematics screeners. This study was guided by the following:

1 How do associations between ASPENS and an end-of-year standardized published measure of mathematics achievement differ between ELs and EPs? What are the associations (i.e., correlations, multiple simultaneous associations) between ASPENS composite, individual subtest, and end-of-year scores for ELs compared to correlations among EPs?
   a *Hypothesis:* Correlations between ASPENS composite and end-of-year scores are predicted to be lower for ELs than for EPs and the extent of variance in end-of-year scores explained by ASPENS performance is predicted to be lower for ELs than for EPs due to the restricted range of scores for students identified as ELs. Further, the association between ASPENS and a comprehensive mathematics test is based on underlying mathematics skills plus error; conversely, the scores for ELs are a function of both mathematics skills and English language skills plus error. Since we do not have a measure of ELs' language skills in the present study, any variation captured by language skill would be modeled as error and make the associations smaller.

2 What are the diagnostic accuracy statistics associated with cut scores based on common definitions to determine risk and severe risk for ELs for ASPENS composite score?

a *Hypothesis:* Cut scores to determine risk on the end-of-year measure for ELs will differ from the cut score for EPs in kindergarten but will be similar in Grade 1 as students have more exposure to both mathematics and English language instruction. As previously described, due to the error created from the differences in administration language and ELs' home languages, we expect the diagnostic statistics to not be as strong for ELs as for EPs.

2. Method

2.1. Participants

Participants were kindergarten and Grade 1 students in six elementary schools from four districts in urban areas of California and Ohio. A total of 715 students, 341 kindergarteners and 374 Grade 1 students. Students participated in ASPENS assessment designed for each respective grade level in the fall and winter, and the TerraNova in the spring. Across grades, 29-33% of each subsample were identified through district criteria as an EL; status as an EL was available for 99% of the sample. EL status was determined via district criteria, which differed across schools, and these procedures included the use of a home language survey or scores on measures of English language development. Assessment data is only available for students who participated in the group administration of the TN and the individual administration of the ASPENS across the fall, winter, and spring timepoints.

2.2. Measures

ASPENS. *Assessing Student Proficiency in Early Number Sense* (ASPENS) is a series of three curriculum-based measures administered for the purposes of universal screening of students' mathematical proficiency. ASPENS measures are designed to assess number sense for both kindergarten and Grade 1 students. Each grade level includes Magnitude Comparison and Missing Number measures, along with an additional subtest designed to assess grade-level specific mathematics content (i.e., Numeral Identification in kindergarten and Base Arithmetic Facts and Base 10 in Grade 1).

Specifically, the kindergarten ASPENS measure includes three subtests: Numeral Identification, Magnitude Comparison, and Missing Number. Numbers within each subtest range from 0 to 20, and the score is the number of correct responses given in one minute. For the Numeral Identification measure, students name numbers as quickly as possible and the Magnitude Comparison measure requires students to name the greater of two visually presented numbers. In the Missing Number measure, students name the missing number in a string of three numbers. Testing for each measure is discontinued after a student misses five consecutive items or one minute has elapsed. Test-retest reliabilities of kindergarten ASPENS measures are in the moderate to high range (.74–.85; Clarke et al., 2011).

The Grade 1 ASPENS measure includes three subtests: Magnitude Comparison, Missing Number, and Basic Arithmetic Facts and Base 10. The first two tests use the same procedures described above, however the range of numbers is 0 to 99. By the end of Grade 1, students should be able to quickly recall some basic arithmetic facts (e.g., 2 + 2 = 4). It is also critical that students retrieve basic arithmetic facts that cross 10 (e.g., 8 + 5; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). To measure these competencies, the Basic Arithmetic Facts and Base 10 measure is added in the middle of Grade 1 (i.e., winter testing time point). Students are presented problems that contain elements that can be composed or decomposed in the Base 10 system (e.g., 5 + 9 becomes 4 + 10) to assess fact fluency. The score is the number of correct items solved in two minutes. Test-retest reliabilities of Grade 1 ASPENS measures also fall within the moderate to high range (.71–.87). Due to the nature of the ASPENS responses (i.e., primarily verbal student responses), the ASPENS was administered individually for each student.

In addition to the individual subtests, there is also a composite ASPENS score for each grade. Raw scores for the subtests are weighted differently and combined, with more weight given to measures that are harder for students at a particular age range (Clarke et al., 2012). An overall ASPENS composite score in kindergarten is calculated by adding the subtest scores as follows: (a) the raw score from Number Identification, (b) the raw score of Magnitude Comparison multiplied by 1.7, and (c) the raw score of Missing Number multiplied by 2.7. For Grade 1, the ASPENS composite score is the sum of: (a) the raw score of Magnitude Comparison multiplied by 1.4, (b) the raw score of Missing Number, and (c) the raw score of Basic Facts multiplied by 1.8.

TerraNova. Mathematics achievement was assessed with the TerraNova, 3rd Edition (TN; CBT/McGraw-Hill, 2011). The TN is a nationally norm-referenced and standardized achievement test used to assess K-12 achievement in reading, language, mathematics, science, and social studies. Form G of the Mathematics subtest was used for the current study; the technical manual reports reliability coefficients at a high of 0.80s for the individual tests (CBT/McGraw-Hill, 2011). The TN is closely aligned with Common Core State Standards, the framework of NAEP, and the National Council of Teachers of Mathematics' most recent standards. Content-related validity for the TN was established via a thorough, nationwide curriculum review. Developers met with educational experts to identify knowledge and skills in each content area and their associated common educational goals. Validation studies were conducted using teacher and student questionnaire to collect data concerning the overall test effectiveness. The reliability of the TN Complete Battery assessment, determined using the Kuder-Richardson Formula 20 coefficient, was .91 (CBT/McGraw-Hill, 2011).

2.3. Procedure

ASPENS was administered to kindergarten and Grade 1 students individually by data collectors with extensive experience in the field. Data collectors, retired teachers with years of experience with whole class assessments, received training on administration of ASPENS and TN during the same school year of test administration, with opportunities to practice. Training was provided to ensure reliable administration of the assessments and that data collectors (i.e., retired teachers) were familiar and comfortable with the testing procedures. They administered the ASPSENS individually on three separate occasions, in November (fall), February (winter), and May (spring), and the TN to student groups in the spring.

2.4. Analytic approach

Prior to additional analyses, descriptive and correlation statistics were evaluated. Regression models determined the extent of variance in the spring TN scores explained by the fall and winter administrations of ASPENS measures. Separate regression models were run for each ASPENS measure and the composite at the fall and winter administrations. All students were included in these analyses, and predictors included the ASPENS measure, a dichotomous indicator of EL status (0 = EP, 1 = EL), and the interaction between these two. All ASPENS measures were mean centered by grade level and spring scores were not used due to the proximity of test administration timepoints for the spring ASPENS measures and the TN. All regression models were run using SPSS 26.

The diagnostic analysis followed the methods outlined in Smolkowski and Cummings (2015) and Cummings and Smolkowski (2015), who provide an accessible introduction to the selection of students at risk for academic difficulties. ASPENS measures across all timepoints and spring TN scores were used in this analysis. Diagnostic statistics were estimated for students at risk, defined by the 35th percentile on the TN. We used the 35th percentile as it aligned with the cut scores used in previous research, with a range from 35th (Gersten et al., 2015) – 40th percentile (Fuchs et al., 2019). We first generated receiver operating characteristic measure (ROC) curves for each of the screening measures administered in the fall, winter, and spring. ROC curves plot the proportion of true positives (sensitivity) against the proportion of false positives (1 – specificity) for all values of the screener. *Positive*, used in these analyses, refers to the indication of risk by the screener; *negative* indicates no risk. Sensitivity reflects the screener's ability to detect risk determined by the criterion (i.e., TN). Specificity reveals the ability of the screener to identify true negatives.

We next calculated the area under the curve, *A*, an overall estimate of the accuracy of the screener. It estimates the average sensitivity (i.e., ability to detect all potential positives) for all levels of specificity (i.e., ability to predict students who really are at some or high risk). Based on a review of the literature on signal detection theory and academic outcomes, Smolkowski and Cummings (2015) considered values of *A* above .95 as excellent, values from .85 to .95 as very good, and values from .75 to .85 as reasonable. Estimates and confidence intervals for *A* were produced by SAS PROC LOGISTIC (SAS Institute, 2016).

We chose decision thresholds based on the screener score with a sensitivity value closest to .80. The decision rule for selecting a decision threshold (i.e., cut score) for each level of risk should depend on the anticipated consequences of each of the four potential outcomes from a screening decision: (1) true positive, (2) false positive, (3) true negative, and (4) false negative (Smolkowski & Cummings, 2015). We selected decision thresholds based on the complement of sensitivity, the proportion of false negatives, such that the cutoff would incorrectly identify no more than 20% of students at risk for mathematics difficulty, based on the TN, as typically achieving (i.e., sensitivity = .80). This decision rule typically produces greater sensitivity than specificity, which would capture fewer false negatives than false positives, which increases the chance that students will receive additional instruction when they need it (i.e., true positives), but also the changes that students will receive additional instruction they do not need (i.e., false positives). This threshold will catch more students who truly require additional support to succeed and the increase in false positives will inflate the number of students who receive supplemental instruction. Due to the added attention these students would receive, however, teachers also have more opportunities to identify students who demonstrate adequate achievement and subsequently exclude them from unnecessary small-group instruction. In contrast, we choose a threshold that minimized false positives and that would increase false negatives, or students who require supplemental support but failed to receive it. These students could then flounder in regular instruction until the next administration of a universal screener. We believe false negative assignments could induce more harm and last longer, as these students may receive only core instruction, then false positives, students who receive more attention in supplemental instruction, where teachers are more likely to detect and remedy decision errors.

For each measure, we reported *A* with confidence intervals, the selected decision threshold (i.e., recommended cut point), sensitivity and specificity with confidence bounds, negative and positive predictive values, the proportion of students who screened positive (τ), and the base rate or proportion determined to be at risk on the criterion measure. Confidence bounds around sensitivity and specificity were formed using a normal-curve approximation (Harper & Reeves, 1999), which are recommended only when cell sizes (e.g., number of false positives, number of true negatives) were greater than 10. We also defined confidence bounds around the cut scores, which represented the lowest and highest screener scores for which the sensitivity confidence intervals contained sensitivity of .80. Frequency statistics were produced with SAS PROC FREQ (SAS Institute, 2016). The online supplemental appendix presents the results for additional combinations of assumptions: risk level defined by the 15th and 35th percentiles with decision thresholds based on sensitivity values of .80 or .90, and the TN criterion test collected at the end of the year following screener administration for risk defined by the 35th percentile.

3. Results

3.1. Descriptive statistics and correlations

Descriptive statistics for all study measures by EL student status are displayed in Table 1. Descriptively, EP kindergarten and Grade 1 students scored slightly higher than ELs on ASPENS at most time points and TN.

Correlations. To address research question 1, correlations for ASPENS composite scores and TN-3 scores across the year are reported in Table 2. For ELs at both grade levels, the fall, winter, and spring ASPENS composite scores were not as strongly correlated with one another and with TN scores compared to the EP group. This pattern was especially pronounced in the correlation between the fall kindergarten ASPENS composite and the TN scores, where the correlation for EL students was $r = .25$ compared to $r = .58$ for EP students.

Multiple Regression. Results for the regression models with spring TN scores regressed on each measure, EL status, and the interaction between the measure and EL status are reported in Tables 3 and 4. Separate regression models were conducted for each measure and the composite score at fall and winter timepoints for kindergarten and Grade 1. For all regression analyses, Benjamini and Hochberg (1995) corrections were used to account for multiple tests and *p* values for the overall regression models reflect this correction procedure. For illustrative purposes, only kindergarten Number Identification will be reviewed. All other kindergarten and all Grade 1 results can be found in Tables 3 and 4, respectively.

For kindergarten Number Identification, approximately 24% of the variance in TN score was explained by the predictors (Number Identification score, EL status, and the interaction between EL status and Number Identification score), $R^2 = .24$, $F(3, 285) = 29.53$, original $p < .001$, adjusted $p < .001$. Further, as we mean-centered ASPENS scores, a student with an average score on fall Number Identification and identified as an EP would be expected to score approximately 462 on the TN (see Table 3, row 1). Students identified as ELs would be expected to score about 7 points less than their EP peers on the TN based on an average score on Number Identification. For a one-point gain on Number Identification, an EP student would be expected to gain 1.37 points on the TN, but ELs would be expected to gain about 1 point less than EPs, $b = -0.99$, $t(1, 283) = -3.40$, $SE = 0.29$, $p = .001$. Further, to determine the expected gain for EL students, the regression coefficients for the expected gain for EP students is subtracted by the interaction coefficient. Therefore, for a one-point gain on Number Identification, EL students would be expected to gain 0.38 points on the TN, $b = 0.38$, $t(1, 285) = 1.51$, $SE = 0.25$, $p = .133$, though this was not statistically significant. While ELs at the grade-level mean score about 7 points higher on the TN than EPs, the association between the Number Identification and TN is weaker; each point gained on Number Identification represents a 1.4 gain in TN scores for EPs but only a 0.4 gain for ELs. For the remaining measures in kindergarten, see Table 3.

Overall, in Grade 1, the interaction terms are consistently not statistically significant, indicating that we could not detect any variation by status in the association between the screeners and spring TN, as depicted in Table 4. That is, if there are two Grade 1 students, one EL and one EP, we cannot conclude that there would be differences in their TN score given they had the same ASPENS composite score. Conversely, if these two students were kindergarten students, the same prediction would not hold true based on EL status. This demonstrates that there may be something being captured in kindergarten outside of the ASPENS measure but related to EL status and impacting student achievement at the end of the school year. These results may indicate that the ASPENS measures may be used with the same procedures and cut off scores in Grade 1, but different screening procedures (e.g., measures, cut scores, directions, etc.) may be needed in kindergarten. The ROC curve analyses provide further detail on if this finding holds true.

Table 1
ASPENS and TerraNova Scores for Kindergarten and Grade 1 by English Learner Status

	Kindergarten			Grade 1		
Statistic	All students	English learners	English proficient students	All students	English learners	English proficient students
ASPENS Fall Composite						
N	317	74	240	349	102	245
Minimum	0	0	0	0	0	0
Maximum	182.3	182.3	168.4	63.8	49.8	63.8
Mean (*SD*)	49.2 (42.2)	49.7 (41.2)	49.0 (41.2)	24.0 (14.4)	22.8 (12.9)	24.7 (14.9)
ASPENS Winter Composite						
N	338	95	240	370	122	245
Minimum	0	0	0	0	0	0
Maximum	211.2	166.8	211.2	141.2	141.2	98.2
Mean (*SD*)	75.2 (43.3)	69.7 (40.4)	78.0 (44.2)	39.2 (20.3)	35.3 (20.0)	41.2 (20.3)
ASPENS Spring Composite						
N	343	106	234	374	124	247
Minimum	8.7	8.7	11.4	0	5.8	0
Maximum	225.4	202.6	225.4	157.8	157.8	126.4
Mean (*SD*)	108.4 (47.4)	104.4 (44.3)	110.8 (48.6)	52.6 (23.0)	51.4 (23.6)	53.4 (22.8)
TerraNova						
N	342	105	233	374	124	246
Minimum	290	405	290	390	404	390
Maximum	623	577	623	680	595	680
Mean (*SD*)	463.0 (40.9)	465.2 (33.3)	462.1 (44.0)	518.5 (42.2)	507.6 (34.8)	524.3 (44.8)

Table 2
Correlations for ASPENS and TerraNova Scores for Kindergarten and Grade 1 Samples by English Learner Status

	ASPENS Composite Scores					
	Kindergarten			Grade 1		
	Fall	Winter	Spring	Fall	Winter	Spring
APSENS Composite Scores						
Winter						
All students	.84			.84		
English learners	.80			.74		
English proficient students	.86			.87		
Spring						
All students	.76	.85		.77	.87	
English learners	.70	.80		.72	.87	
English proficient students	.78	.86		.80	.87	
TerraNova						
All students	.50	.53	.56	.57	.63	.63
English learners	.25	.37	.50	.53	.67	.72
English proficient students	.58	.590	.60	.57	.61	.61

Note. Fall indicates scores at the fall testing timepoint, Winter indicates scores for the mid-year testing time point, and Spring indicates scores obtained at the end-of-year testing timepoint. TerraNova scores are standard scores. All correlations were statistically significant at $p < .05$.

Table 3
Results of Regression of ASPENS Subtests, English Learner Status, and Their Interaction on Spring TerraNova for Kindergarten Students

			Predictors								
Measure	R^2	*Intercept* (*SE*)	Measure			EL status			Measure x EL status		
			b^*	*b*	*SE*	b^*	*b*	*SE*	b^*	*b*	*SE*
Fall											
Number Identification	.237	461.58 (2.45)	.56	1.37***	0.15	.07	7.08	4.93	-.20	-0.99**	0.29
Magnitude Comparison	.221	462.85 (2.47)	.52	2.64***	0.30	.06	5.2	4.97	-.17	-1.86**	0.64
Missing Number	.263	463.90 (2.33)	.57	4.00***	0.57	.04	3.20	4.71	-.15	-2.20**	0.85
Composite	.292	463.09 (2.29)	.61	0.58***	0.06	.05	5.01	4.60	-.20	-0.38**	0.11
Winter											
Number Identification	.223	459.61 (2.45)	.51	1.30***	0.14	.11	10.23*	4.58	-.11	-0.55	0.30
Magnitude Comparison	.275	459.85 (2.36)	.58	2.50*	0.24	.11	9.67*	4.40	-.15	-1.28**	0.48
Missing Number	.284	461.55 (2.34)	.59	4.39***	0.41	.06	5.67	4.32	-.15	-2.12**	0.80
Composite	.314	459.91 (2.30)	.62	0.59***	0.05	.10	9.46*	4.27	-.15	-0.29**	0.10

Note. b^* = standardized regression coefficients. *b* = unstandardized regression coefficients. For all F tests, $p < .001$ (Benjamini–Hochberg-adjusted), tested with 3 and 284–285 *df* in the fall and 3 and 316–317 *df* in the spring. EL status was a dichotomous predictor (0 = English proficient students, 1 = English learners). ASPENS measures were centered on their mean to aid interpretation.

*unadjusted $p < .05$.

**unadjusted $p < .01$.

***unadjusted $p < .001$.

Table 4
Results of Regression of ASPENS Subtests, English Learner Status, and Their Interaction on Spring TerraNova for Grade 1 Students

Measure	R^2	*Intercept* (*SE*)	Predictors								
			Measure			EL status			Measure x EL status		
			*b**	*b*	*SE*	*b**	*b*	*SE*	*b**	*b*	*SE*
Fall											
Magnitude Comparison	.295	521.79 (2.30)	.56	2.92***	0.28	-.05	-4.26	4.31	-.05	-0.58	0.59
Missing Number	.296	523.02 (2.29)	.58	4.53***	0.43	-.09	-8.26	4.24	-.10	-1.53	0.82
Composite	.326	522.23 (2.24)	.60	1.69***	0.15	-.07	-5.89	4.18	-.07	-0.42	0.30
Winter											
Base 10	.277	523.00 (2.32)	.56	6.02***	0.63	-.12	-10.90**	4.04	-.10	-1.81	1.01
Magnitude Comparison	.381	522.53 (2.15)	.63	3.32***	0.27	-.10	-8.56*	3.77	-.07	-0.61	0.48
Missing Number	.313	523.03 (2.26)	.55	3.81***	0.37	-.11	-9.47*	3.96	-.03	-0.35	0.67
Composite	.409	514.21 (3.02)	.66	1.33***	0.10	-.09	-8.00*	3.68	-.07	-0.23	0.18

Note. *b* = unstandardized regression coefficients. *b** = standardized regression coefficients. EL status was a dichotomous predictor (0 = English proficient students, 1 = English learners). Benjamini-Hochberg correction used to account for multiple tests. ASPENS measures were centered on their mean to aid interpretation.
*unadjusted $p < .05$.
**unadjusted $p < .01$.
***unadjusted $p < .001$.

Table 5
Diagnostic Statistics and Cut Scores for ASPENS Screeners Collected in Kindergarten Compared to the Criterion 35th Percentile on the Spring Administration of the TerraNova in the Same School Year at Target Sensitivity of .80

	A		Cut Points [a]		Observed Sensitivity (TPF)		Specificity (1 – FPF)		NPV	PPV	Predicted Positive	Base Rate
Fall Assessments												
NI	.719	[.659, .779]	28	[22, 31]	.803	[.731, .875]	.416	[.343, .489]	.758	.482	.672	.403
MC	.745	[.688, .802]	7	[6, 10]	.786	[.712, .860]	.532	[.458, .606]	.786	.532	.597	.403
MN	.779	[.725, .833]	**7**	[5, 8]	.819	[.749, .889]	.572	[.498, .646]	.825	.562	.585	.401
Composite	.773	[.718, .828]	**50**	[42, 64]	.802	[.729, .875]	.607	[.534, .680]	.820	.578	.557	.401
Winter Assessments												
NI	.720	[.663, .777]	38	[35, 41]	.805	[.738, .872]	.487	[.416, .558]	.780	.525	.634	.413
MC	.770	[.719, .821]	**15**	[14, 16]	.811	[.744, .878]	.608	[.538, .678]	.821	.591	.564	.411
MN	.809	[.760, .858]	**8**	[7, 9]	.805	[.738, .872]	.656	[.588, .724]	.827	.622	.534	.413
Composite	.789	[.739, .839]	**82**	[70, 89]	.803	[.735, .871]	.619	[.550, .688]	.818	.596	.555	.411
Spring Assessments												
NI	.722	[.668, .776]	46	[44, 48]	.797	[.731, .863]	.508	[.438, .578]	.775	.540	.621	.421
MC	.789	[.741, .837]	**22**	[20, 23]	.811	[.747, .875]	.604	[.536, .672]	.815	.598	.571	.421
MN	.740	[.687, .793]	16	[14, 16]	.825	[.763, .887]	.492	[.422, .562]	.795	.541	.641	.421
Composite	.778	[.729, .827]	**116**	[106, 125]	.804	[.739, .869]	.584	[.515, .653]	.804	.584	.579	.421

Note. NI = numeral identification, MC = magnitude comparison, MN = missing number, *A* = area under the ROC curve, TPF = true-positive fraction, FPF = false-positive fraction, NPV = negative predictive value, and PPV = positive predictive value. Decision thresholds (cut points) bolded if AUC ≥ .75. The 95% confidence intervals were provided for AUC, sensitivity, and specificity. A dash (–) indicates unreliable confidence intervals due to cell sizes with 10 or fewer cases. The base rate is the observed proportion of students who scored below the 35th percentile.
[a] Bounds on cut points represent the lowest and highest cut points for which the sensitivity confidence interval contains the target sensitivity value.

3.2. Diagnostic accuracy of ASPENS

Regarding research question 2, Tables 5 and 6 report the results of the diagnostic analyses and cut score selection for all kindergarten and Grade 1 measures using the 35th percentile on the TN collected in the spring of the same year. Each decision threshold was based on the a priori decision to maintain a sensitivity level closest to .80. Table 5 reports results for the kindergarten ASPENS measures in detail for students likely to possess mathematics difficulties at the end of the year and Table 6 reports the results for the Grade 1 ASPENS measures. We examined the difference between screener performance between EL and EP students with a series of ROC curves. Due to the number of estimates in Tables 5 and 6, we first provide an example of their interpretation.

3.3. Example: ASPENS composite in Kindergarten

We defined students with mathematics difficulties as those falling below the 35th percentile on the TN in the spring of kindergarten. All others were considered not at risk for estimates of diagnostic statistics. The screener data, therefore, fall into two overlapping distributions. One distribution describes students who demonstrate mathematics difficulties at the end of the year. For illustration, the fall composite distri-

Table 6
Diagnostic Statistics and Cut Scores for ASPENS Screeners Collected in Grade 1 Compared to the Criterion 35th Percentile on the Spring Administration of the TerraNova in the Same School Year at Target Sensitivity of .80

	A		Cut Points [a]		Observed Sensitivity (TPF)		Specificity (1 – FPF)		NPV	PPV	Predicted Positive	Base Rate
Fall Assessments												
MC	.781	[.730, .832]	**15**	[14, 16]	.802	[.734, .870]	.616	[.548, .684]	.824	.580	.550	.398
MN	.777	[.726, .828]	**10**	[9, 10]	.817	[.751, .883]	.556	[.487, .625]	.821	.549	.593	.398
Composite	.793	[.744, .842]	**26**	[24, 28]	.809	[.742, .876]	.646	[.579, .713]	.837	.602	.535	.398
Winter Assessments												
MC	.812	[.765, .859]	**18**	[16, 19]	.785	[.719, .851]	.700	[.638, .762]	.819	.654	.503	.419
MN	.803	[.757, .849]	**12**	[11, 12]	.826	[.765, .887]	.643	[.578, .708]	.836	.624	.553	.419
BFB10	.809	[.764, .854]	**4**	[4, 4]	.785	[.719, .851]	.696	[.633, .759]	.818	.650	.506	.419
Composite	.841	[.798, .884]	**37**	[35, 42]	.805	[.741, .869]	.783	[.727, .839]	.848	.727	.463	.419
Spring Assessments												
MC	.816	[.772, .860]	**23**	[22, 24]	.791	[.728, .854]	.674	[.611, .737]	.815	.641	.523	.424
MN	.781	[.734, .828]	**14**	[13, 15]	.804	[.742, .866]	.558	[.492, .624]	.795	.572	.595	.424
BFB10	.771	[.722, .820]	**9**	[8, 10]	.804	[.742, .866]	.521	[.454, .588]	.783	.552	.617	.424
Composite	.830	[.788, .872]	**55**	[50, 58]	.797	[.734, .860]	.660	[.597, .723]	.816	.633	.534	.424

Note. MC = magnitude comparison, MN = missing number, BFB10 = basic facts and base 10, *A* = area under the ROC curve, TPF = true-positive fraction, FPF = false-positive fraction, NPV = negative predictive value, and PPV = positive predictive value. Decision thresholds (cut points) bolded if AUC ≥ .75. The 95% confidence intervals were provided for AUC, sensitivity, and specificity. A dash (–) indicates unreliable confidence intervals due to cell sizes with 10 or fewer cases. The base rate is the observed proportion of students who scored below the 35th percentile.

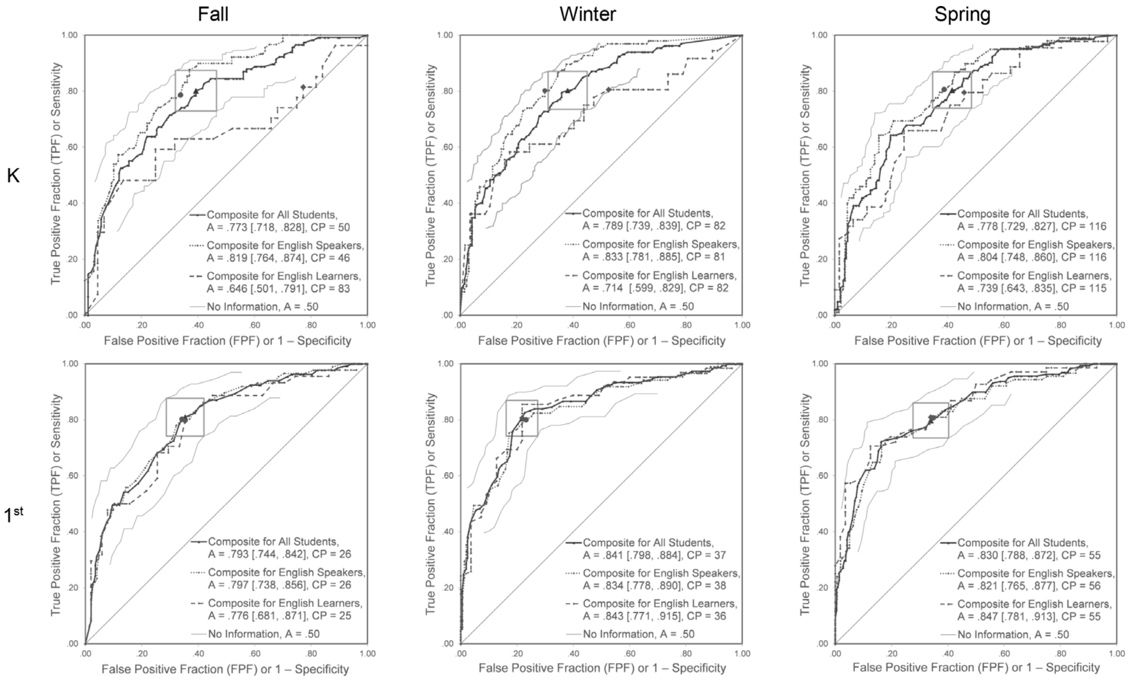

Fig. 1. ROC curves for the composite score on ASPENS used to discriminate students with some risk of math difficulties, determined by the 35th percentile on the TerraNova. The dark solid line depicts the ROC curve for all students, the light lines the confidence bounds around the curve, and the box show the confidence bounds around the chosen cut score. The dotted lines show the ROC curve for students proficient in English and the dashed lines for the English learners. The large markers show the chosen cut score defined as the point with sensitivity closest to .80. The legend in each figure provides the area under the ROC curve, *A*, and the chosen cut point (CP). Notably, except for the fall of kindergarten, our decision rules selected a similar cut score for each sample.

bution for students with mathematics difficulties has a mean of 28.7 with a standard deviation of 30.2. The other curve describes students not at risk, *M* = 64.8, *SD* = 42.0. The ROC curve (see Fig. 1) describes the separation between the two distributions. The *A* metric quantifies the separation between the two distributions by reporting the average sensitivity over all values of specificity. With the criterion, ASPENS fall kindergarten composite demonstrated moderate predictive accuracy for the fall, winter, and spring with areas under the ROC curves, *A*, of .77, .79, and .78, respectively.

We next selected decision thresholds based on sensitivity. Students with mathematics difficulties determined by the TN had an 80% chance of being identified as such with a cut score of 50 in the fall, 82 in the winter, and 116 in the spring. The observed sensitivity at each of these selected cut scores were between .802 and .804 (see Table 5). Specificity describes the proportion of students considered not at risk by the screener who are not at risk according to the criterion set by the authors. With decision thresholds set at sensitivity of .80 for kindergarten, 61%, 62%, and 58% of students were correctly identified as not at risk in the fall, winter, and spring.

Predictive values indicate the clinical significance of a screener (Pepe, 2003). For the cut score of 50 on the fall composite, the positive predictive value (PPV) indicates that slightly over half (58%) of the students who screened positive scored below the 35th percentile on the TN. The negative predictive value (NPV) shows that of those students who screened not at risk, 82% scored where we predicted they would be, at or above the 35th percentile on the spring TN. It should be noted that predictive values depend on the proportion of students with mathematics difficulties (i.e., the *base rate*). Districts with a lower prevalence of mathematics difficulties (i.e., lower base rate) will experience different predictive values.

3.4. Overall accuracy of ASPENS screeners

Tables 5 and 6 present results for all ASPENS measures in kindergarten and Grade 1, with the 35th percentile on the end of year TN set as criterion for defining mathematics difficulty. The areas under the ROC curves, A, were moderate, from .72 to .84, with A increasing as students progress from the fall of kindergarten to the spring of Grade 1. The composite performed well for all assessment times, but some subscales, such as Number Identification in kindergarten, did not perform adequately (i.e., $A < .75$) in the present sample. Overall, the results suggest that ASPENS can discriminate students' risk for mathematics difficulties and the value of the screener improved over time.

ELs and EPs. To examine differences between ELs and EPs, we plotted ROC curves for the ASPENS composite for ELs, EPs, and the full sample (see Fig. 1). Each figure shows the decision threshold (i.e., cut point) chosen for the composite screener with a large marker, and for the full sample, we plotted a box surrounding the decision threshold to show the confidence bounds on sensitivity and specificity. The plots also show the confidence bounds around the full sample, represented by thin, solid, lines. The legend in each figure provides the area under the ROC curve, A, and the chosen cut point (CP) for each sample.

The ROC curves demonstrate that, although the screener performs differently for ELs and EPs in the fall of kindergarten, these differences converge through kindergarten. By Grade 1, the three samples essentially overlap. In addition, except for the fall of kindergarten, our decision rules selected a similar cut score for each sample. The cut scores for ELs and EPs from the spring of kindergarten through Grade 1 fall within the confidence bounds for the full sample. Moreover, even though the specificity values differ for the three groups in the winter of kindergarten, the cut scores were nearly identical (e.g., 81 or 82).

4. Discussion

The purpose of this investigation was to analyze the psychometric properties of ASPENS for ELs in kindergarten and Grade 1 and to determine whether these screening measures identify risk status for EL and EP students similarly. Our study included three primary objectives: (a) examine the descriptive statistics for EPs compared to ELs on the ASPENS measures, (b) determine the differences for ELs and EPs in the extent of variance explained in students' mathematics outcome scores by ASPENS scores, and (c) examine diagnostic accuracy statistics to determine risk on a mathematics outcome measures for ELs and EPs. Given calls to investigate early numeracy screening measures for ELs (Purpura et al., 2015), this study represents a first contribution to the research base, providing initial insight into how a widely used early numeracy screener functions for EL and EP students. This investigation provides evidence to the early childhood education field regarding the common practice of using the same screener for all students, regardless of EL status.

4.1. ASPENS performance by EL status

Our descriptive analyses indicated that in general, ELs had lower composite ASPENS scores across testing time points compared to EPs. Correlations among the composite scores across time points were slightly higher for EPs than ELs. The correlations between ASPENS composite and TN were relatively stable for EPs, ranging from .57 to .61 from the fall of kindergarten to spring of Grade 1. For ELs, however, the correlations with the TN increase over time, from .25 in the fall of kindergarten to .72 in the spring of Grade 1. For ELs, the lower correlations in kindergarten may have been a result of their limited English proficiency.

For measures of mathematics, English language proficiency could be interpreted as construct-irrelevant variance (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), or variability attributed to skills or knowledge that is irrelevant to the test's purpose. Most instruction in U.S. schools, however, takes place in English and both ASPENS and TN assess mathematics knowledge and skill in English. Though it may be tempting to attribute the correlation differences in kindergarten as predictive bias, as Cummings et al. (2021) point out, predictive bias assumes the grouping variable, such as EP versus EL here, is inconsequential to performance. Cummings et al. (2021) discuss several papers on reading screeners that draw this conclusion. With mathematics measures, English language is consequential, so the correlation differences would not represent predictive bias. Importantly, instruction that teaches mathematics vocabulary in English explicitly can support ELs with limited English language proficiency (e.g., Doabler et al., 2019). Hence, it is likely that proficiency with English among ELs improved over time as students received instruction in both the English language and mathematics in English, which would then account for the improved correlation from the fall of kindergarten to Grade. 1.

In the regression analyses for our first research question, we examined the extent to which student scores on the ASPENS composite and individual subtest explained variance on the TN, including EL status as a predictor. The regression results demonstrated that ELs consistently scored lower on the ASPENS measure compared to their EP peers across both grade levels. Results that approached statistical significance were found more often in kindergarten than in Grade 1. These results indicate that there may be a change in the relationship between EL status and performance on the measures over time. The amount of variance accounted for by students' ASPENS performance and EL status is higher when the student is in Grade 1 than kindergarten. For instance, the composite score accounted for approximately 30% of the variance ($R^2 = .292$) in fall of kindergarten, but 33% of variance in TN score in fall of Grade 1 ($R^2 = .326$). Larger differences between EPs and ELs were evident in fall testing in kindergarten than in Grade 1, suggesting that there is either maturation, intervention, increased English language proficiency for ELs, or some other combination of effects that reduce group differences by the next school year. Later screening administration may help overcome the impact of these differences. Specifically, we found that it only took one school year for ELs to test similarly to EPs. Further, the relation differing across kindergarten and Grade 1 may demonstrate that the ASPENS measure tasks captured something related to EL status in kindergarten that could be impacting student achievement at the end of the school year. These differences may also be due to the ASPENS measure not identifying the most predictive skills early in kindergarten for ELs and the current instructional paradigm, English-only mathematics instruction, is failing these students.

Answering research question 2, the ROC analyses demonstrated that the measures could discriminate students' risk for mathematics difficulties with the screener improving over time. We found that the ASPENS measures identified ELs and EPs similarly in Grade 1. Further, though the screener performed differently for ELs and EPs in the fall of the kindergarten year, these differences diminished across time. There is a need to improve the accuracy of screening, particularly for ELs, in kindergarten based on the research available related to general screening practices and assessing ELs specifically. This investigation demonstrates that using later screening measures, such as ASPENS in winter, may help improve screening accuracy.

4.2. Screening procedures

Overall, our results demonstrate ELs and EPs can be screened using the same measure and cut score with accuracy. Results suggest that beyond the fall of kindergarten, similar scores and associated technical characteristics exist for ELs and EPs. Screeners may perform similarly for students learning English and those proficient in English because language affects both the screeners and criterion measures similarly. ELs are also held to the same standard, at or above the 35th percentile on the TN, as EPs. These findings are promising for schools, as using a single screening measure to determine risk for all students in a grade level is more feasible and cost-effective than differentiating based on EL status. One consideration for schools, though, is that the criterion measure here was also administered in English and therefore may not be the best assessment of ELs' full knowledge or skill in mathematics. Future work should investigate the relation between an English-only measure and an outcome measure in the students' native language. Evaluating the appropriateness of screening measures for all students, including ELs, can assist educators in determining if they are accurately identifying students who struggle. The measures investigated here can be useful for educators, though they should be interpreted with greater caution in fall of kindergarten than other time points (i.e., later in kindergarten and Grade 1).

While our findings indicate that ELs and EPs performed similarly on ASPENS, we caution readers that assessing ELs only in English does not align with best practices. Assessing ELs in their native language continues to be the gold standard but may not be feasible for schools to implement consistently. Our findings demonstrate that, when assessing students in their native language is not available, using English early mathematics screeners may be appropriate.

4.3. Assessment practices for ELs

As we develop a greater understanding of ELs' mathematics development, future research should be attuned to incorporating best practices for ELs. To provide the fullest picture of strengths and areas of need for ELs, assessing students in all languages in which they are fluent is recommended for assessment practices more broadly. Assessment of ELs using inappropriate testing procedures has historically led to ELs being underrepresented for special education services in the elementary grades and overrepresented in fifth grade and beyond (Artiles et al., 2002). Particularly important for ELs, mathematics assessments must adequately provide data that speaks to mathematics skills, not English proficiency (e.g., Cho et al., 2020). With screening practices representing the gateway to more intensive assessment and intervention, it is critical to ensure that screening measures are accurate and that students are classified based on early mathematics challenges, not English proficiency, or lack thereof. ELs are often assessed using the same measures (e.g., screeners, statewide tests, etc.) as EPs (Obuon, 2019), and though it is more feasible for schools to continue this practice, these common assessments are also often not linked with an intervention that is evidence-based for ELs (Doabler et al., 2016).

Our results suggest that a measure of early numeracy may overidentify EL students in need of support, specifically in their kindergarten year. Experts have expressed concern with this possibility more broadly for early numeracy screening measures (Gersten et al., 2009; Gersten et al., 2012). This may result in educators using greater caution when interpreting screening results, using a combination of screening results to officially identify, or waiting until a later date to screen. Our results indicate that screening in the winter may overcome this concern.

The term "EL" is very broad and investigation is needed to unpack how measures might work differently for different populations of ELs. Future work should include a capturing additional data regarding cultural and linguistic backgrounds of students included in a larger sample. This information, along with more descriptive information about the classification of ELs, could assist in determining if these results can be generalized to a different population of ELs and how initial language and mathematics skill impacts the findings. Additionally, increasing the assessment frequency in research can help find the "tipping point" at which EL status no longer impacts sensitivity. Disseminating these results and the results of other investigations regarding differences in performance of ELs and EPs is important for practitioners to make meaningful decisions for their kindergarten students, and potentially beyond.

Our results indicate ASPENS works better as a long-term predictor of student performance than short-term. Specifically, these measures provide critical information at the beginning of the year, particularly in Grade 1. As we learn more about ELs' mathematics development and screening for ELs' risk, we should not lose focus on the importance of early intervention. Integrating intervention into work to develop and validate early screeners will enable the field to investigate factors associated with intervention response. Coupled with more robust understanding of language development should enable schools to better support the learning needs of all students in understanding mathematics.

CRediT authorship contribution statement

Tasia Brafford: Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Ben Clarke:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision. **Russell M. Gersten:** Conceptualization, Investigation, Resources. **Keith Smolkowski:** Methodology, Validation, Formal analysis, Data curation, Writing – review & editing, Visualization, Supervision. **Marah Sutherland:** Writing – original draft, Writing – review & editing. **Joe Dimino:** Conceptualization, Resources. **David Fainstein:** Writing – original draft, Writing – review & editing.

Data Availability

The authors do not have permission to share data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecresq.2022.12.007.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.*

Artiles, A. J., Harry, B., Reschly, D. J., & Chinn, P. C. (2002). Over-identification of students of color in special education: A critical review. *Multicultural Perspectives, 4*, 3–10.

Baker, S., Gersten, R., & Lee, D.-S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *The Elementary School Journal, 103*, 51–73.

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 15*, 775–785. 10.1111/desc.12155.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodology), 57*, 289–300.

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

Bodovski, K., & Farkas, G. (2007). Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction. *The Elementary School Journal, 108*, 115–130. 10.1086/525550.

Booth, J., & Siegler, R. S. (2008). Numerical magnitude representations include artithmatic learning. *Child Development, 79*, 1016–1031. 10.1111/j.1467-8624.2008.01173.x.

Calderón, M., Slavin, R., & Sánchez, M. (2011). Effective instruction for English learners. *The Future of Children, 21*, 103–127. https://www.jstor.org/stable/41229013.

U.S. Department of Education. (2017). 2014–15. SEA File C141. *LEP Enrolled EDFacts Data Warehouse*. https://www2.ed.gov/about/inits/ed/edfacts/data-files/school-status-data.html.

Case, Y., & Okamoto, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development,* 61, 27-58. https://doi.org/10.1111/j.1540-5834.1996.tb00536.x

CBT/McGraw-Hill. (2011). *Terranova, third edition complete battery*. Monterey, CA: CTB/McGraw-Hill.

Chard, D. J., Clarke, B., Baker, S., 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*(2), 3–14. 10.1177/073724770503000202.

Cho, E., Fuchs, L. S., Seethaler, P. M., Fuchs, D., & Compton, D. L. (2020). Dynamic assessment for identifying Spanish-speaking English learners' risk for mathematics disabilities: Does language of administration matter? *Journal of Learning Disabilities, 53*, 380–398. 10.1177/0022219419898887.

Clarke, D., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). *A randomized control trial of a Tier 2 kindergarten mathematics intervention.*

Clarke, B., Doabler, C. T., & Nelson, N. J. (2014). Best practices in mathematics assessment and intervention with elementary students. P. Harrison & A. Thomas. *Best practices in school psychology: Data-based and collaborative decision making, 1*, 219–232 6th ed.,.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (ASPENS)*. Sopris Learning: Cambium Learning Group.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234–248.

Cummings, K. D., & Smolkowski, K. (2015). Bridging the gap: Selecting students at risk for academic difficulties. *Assessment for Effective Intervention*, 41(1), 55-61. https://doi.org/10.1177/1534508415590396

Cummings, K. D., Smolkowski, K., & Baker, D. L. (2021). Comparison of literacy screener risk selection between English proficient students and English learners. *Learning Disability Quarterly, 44*(2), 96–109. 10.1177/0731948719864408.

Doabler, C. T., Clarke, B., Kosty, D., Smolkowski, K., Kurtz-Nelson, E., Fien, H., & Baker, S. K. (2019). Building number sense among English learners: A multisite randomized controlled trial of a Tier 2 kindergarten mathematics intervention. *Early Childhood Research Quarterly, 47*, 432–444. http://doi.org/10.1016/j.ecresq.2018.08.004.

Doabler, C. T., Nelson, N. J., & Clarke, B. (2016). Adapting evidence-based practices to meet the needs of English learners with mathematics difficulties. *Teaching Exceptional Children, 48*(6), 301–310.

Driver, M. K., & Powell, S. R. (2017). Culturally and linguistically responsive schema instruction: Improving word problem solving for English language learners with mathematics difficulties. *Learning Disability Quarterly, 40*, 41–53. 10.1177/0731948716646730.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. 10.1037/0012-1649.43.6.1428.

Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., Hamlett, C. L., & Zumeta, R. O. (2009). Remediating number combination and word problem deficits among students with mathematics difficulties: A randomized control trial. *Journal of Educational Psychology, 101*, 561–576. 10.1037/a0014701.

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*, 311–330. 10.1177/001440290707300303.

Fuchs, L. S., Fuchs, D., & Gilbert, J. K. (2019). Does the severity of students' preintervention math deficits affect responsiveness to generally effective first-grade intervention? *Exceptional Children, 85*, 147–162. 10.1177/0014402918782628.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools. NCEE 2009-4060. What Works Clearinghouse.

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

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized control trial. *American Educational Research Journal, 52*, 516–546. 10.3102/0002831214565787.

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*, 118–126. 10.1177/1534508411414151.

Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology, 93*, 615–627. 10.1037/0022-0663.93.3.615.

Harper, R., & Reeves, B. (1999). Reporting of precision of estimates for diagnostic accuracy: A review. *BMJ : British Medical Journal, 318*(7194), 1322–1323. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC27871/ .

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*, 181–195.

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22*, 36–46.

La Paro, K. M., Hamre, B. K., Locasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D., Clifford, R., Barbarin, O., Howes, C., & Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development,* 20, 657–692. https://doi.org/10.1080/10409280802541965

Lembke, E. S., Foegen, A., Whittaker, T. A., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention, 33*(4), 206–214. 10.1177/1534508407313479.

Mazzocco, M. M. (2005). Challenges in identifying target skills for math disability screening and intervention. *Journal of Learning Disabilities, 38*, 318–323. 10.1177/00222194050380040701.

Morgan, P. L., Hillemeier, M. M., Farkas, G., & Maczuga, S. (2014). Racial/ethnic disparities in ADHD diagnosis by kindergarten entry. *Journal of Child Psychology and Psychiatry, 55*, 905–913. 10.1111/jcpp.12204.

National Assessment of Educational Progress. (2019). *The nation's report card. Mathematics 2019: National Assessment of Educational Progress at grades 4 and 8*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education https://www.nationsreportcard.gov/mathematics/?grade=4.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards mathematics* http://www.corestandards.org/.

Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology, 111*(6), 1001–1022. 10.1037/edu0000334.

Obuon, O. P. (2019). Assessing Mathematics or Language Proficiency? *Relationship between english language proficiency and mathematics achievement among English learners*. Doane University [Dissertation].

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford.

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology, 110*, 647–658. https://doi.org/10.1016/j.jecp.2011.07.004.

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review, 44*, 41–59. 10.17105/SPR44-1.41-59.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853–891. 10.3102/0002831209333184.

Richards-Tutor, C., Baker, D. L., Gersten, R., Baker, S. K., & Smith, J. L. M. (2016). The effectiveness of reading interventions for English learners: A research synthesis. *Exceptional Children, 82*, 144–169. 10.1177/0014402915585483.

SAS Institute. (2016). *Base SAS® 9.4 procedures guide: Statistical procedures*. SAS Institute.

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children, 77*, 7–59. 10.1177/001440291007700102.

Smolkowski, K., & Cummings, K. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention, 41*(1), 41–54. 10.1177/1534508415590386.

U.S. Department of Education, National Center for Education Statistics, Common Core of Data (2017). "State Nonfiscal Survey of Public Elementary/Secondary Education," 2014–15. https://nces.ed.gov/ccd/stnfis.asp

Thomas, W. P., & Collier, V. P. (2002). A national study of school effectiveness for language minority students' long-term academic achievement. https://files.eric.ed.gov/fulltext/ED475048.pdf

U.S. Department of Education, Office of English Language Acquisition, National Clearinghouse for English Language Acquisition. (2017). *Profiles of English Learners (ELs)*. http://www.ncela.us/files/fast_facts/05-19-2017/ProfilesOfELs_FastFacts.pdf