Exceptional Children

Vol. 78, No. 4, pp. 423-445.
©2012 Council for Exceptional Children.

Universal Screening in Mathematics for the Primary Grades: Beginnings of a Research Base

RUSSELL GERSTEN
Instructional Research Group

BEN CLARKE
University of Oregon

NANCY C. JORDAN
University of Delaware

REBECCA NEWMAN-GONCHAR

KELLY HAYMOND
Instructional Research Group

CHUCK WILKINS
Edvance Research, Inc.

ABSTRACT: *This article describes key findings from contemporary research on screening for early primary grade students in the area of mathematics. Existing studies were used to illustrate the constructs most worth measuring and the diverse strategies that researchers used to study potential measures. The authors discussed the strengths and weaknesses of assessing a few key proficiencies (as is often done in early reading) versus a more full-scale battery, and described the importance of going beyond merely reporting predictive validity correlation coefficients to examining the classification accuracy, specificity, and sensitivity of screening measures.*

Recent longitudinal research strongly suggests that students who perform poorly on simple mathematics problems at the end of kindergarten and first grade are likely to continue to perform poorly in mathematics through fourth grade (Duncan et al., 2007; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Morgan, Farkas, & Wu, 2009). In fact, using a nationally representative sample of students, Morgan et al. (2009) found that students who remained in the lowest 10th percentile at both the beginning and end of kindergarten (often considered an indicator of a learning disability in mathematics) had a 70% chance of remaining in the lowest 10th percentile

5 years later. They also tended to score, on average, two standard deviation units (48 percentile points) below students who were in the acceptable range of mathematics performance in kindergarten. Jordan et al. (2009) found that kindergartners' number sense, that is, their knowledge of number relationships and the meaning of number concepts, predicts later mathematics achievement even when statistically controlling for IQ and socioeconomic status.

Just as the persistence of reading disabilities stimulated widespread investment in early intervention and screening in reading, we hope that the concurrent findings for the persistence of mathematics difficulties will incite similar leaps for identifying measures to screen students likely to experience difficulties in mathematics. In the past decade, a number of mathematics screening measures for use in the primary grades have been developed. We can draw some reliable conclusions from the convergent findings of these efforts. Using extant studies of early screening in mathematics, we illustrate principles and draw attention to issues to help professionals understand what mathematics constructs measure as well as strengths and weaknesses of contemporary screening measures.

Additionally, we examine a critical, but rarely explored, issue in research on screening in education: *classification accuracy,* that is, the precision with which measures accurately detect which students will have trouble in mathematics without intensive intervention. Because this area of research is underdeveloped in mathematics, this article will describe the concept of classification accuracy, specifically sensitivity and specificity, and their relationship to decisions educators must make when selecting a screening measure.

APPROACHES TAKEN TOWARDS MEASURING NUMBER PROFICIENCY IN YOUNG STUDENTS

Virtually all math screening measures for the primary grades rely on assessing aspects of what is often referred to as *number sense.* Okamoto and Case (1996) describe number sense as the development of increasingly sophisticated understanding of numbers and understanding that is typically represented by students' ability to use increasingly sophisticated mental number lines. Individuals with good number sense appear to develop a mental number line on which they can represent and manipulate numerical quantities. However, number sense is more complex than the development of a mental number line. Berch (2005) captures the complexities of articulating a working definition of number sense, remarking, "Possessing number sense ostensibly permits one to achieve everything from understanding the meaning of numbers to developing strategies for solving complex math problems; from making simple magnitude comparisons to inventing procedures for conducting numerical operations" (p. 334). For that reason, the National Research Council (2009) recommended use of the term *number proficiencies* to refer to the specific components of number sense that are the focus of an assessment or an intervention. Because both terms have been used historically, we use both in this article.

Researchers have adopted an array of approaches for developing assessments of early number proficiency/number sense. The first approach attempts to develop efficient screening measures so that schools can discern which students are likely to require additional assistance. Many of these researchers (e.g., Lembke & Foegen, 2009) have focused on the development of a set of brief, timed measures, each of which gauges one key aspect of number competence. Typically, the research team calculates predictive validity indices for each individual test. Recently, researchers have used some form of ordinary least squares regression to develop a composite score (often the sum or average of each individual measure) that predicts subsequent mathematics achievement. The second approach for developing screening measures includes the development of one measure of number proficiency that intentionally samples across several different aspects of number proficiency. Examples are the research of Bryant, Bryant, Gersten, Scammacca, and Chavez (2008) and the recent research of Jordan, Glutting, and Ramineni (2010).

Procedure for Reviewing the Literature on Early Screening in Mathematics

We conducted a literature review using the ERIC and PsycINFO databases. We used the descriptors *screening* and *mathematics* and limited our search to empirical studies published between 1996 and 2011. We limited our search to studies involving children ranging in age from birth to 12 years old and excluded dissertations. We also conducted a manual search of major journals in special, remedial, and elementary education (*Journal of Special Education*, *Exceptional Children*, *Journal of Educational Psychology*, and *Journal of Learning Disabilities*) to locate relevant studies.

This search resulted in the identification of 48 studies. Of this total, 21 studies were selected for further review based on analysis of the title, abstract, and keywords. Of these 21 studies, 16 (76%) met our criteria for inclusion. Out of the 16 studies identified, eleven focused on single proficiency measures, four included multiple proficiency measures, and five used diagnostic utility statistics and receiver operating characteristics (ROC) analyses to predict mathematics learning disability (MLD) or low-achieving students. (Seethaler and Fuchs, 2010, used a single proficiency measure, a multiple proficiency measure, and diagnostic utility statistics. Geary, Bailey, and Hoard, 2009, used a single proficiency measure and diagnostic utility statistics. Clarke et al., 2011, used a multiple proficiency measure and diagnostic utility statistics.)

Our criteria for inclusion limited our review to studies that targeted kindergarten and first grade students, included screening measures and outcome variables specific to mathematics performance, and reported predictive validity, ROC curves, or sensitivity and specificity analyses. Two of the studies, Locuniak and Jordan (2008) and Bryant et al. (2008), included both a first- and a second-grade sample. Although we limit the data presented in the tables to first grade only, we discuss the second-grade sample in the text of this article. We excluded studies that used one or more norm-referenced standardized measures as a screener because we were interested in an efficient screener or screening batteries. Many of the standardized measures are much longer than we would recommend for a screener, often taking between 1 hr and 3 hr.

For the data presented in Tables 1 and 2, we focused on studies that provided correlations between screeners administered in the fall and mathematics outcomes administered in the spring of that same year (in one case, we included a screener given in the spring of the preceding year since this is a practice that some districts use). For the tables specifically, we compared measures across a similar time frame because most schools screen students in the fall as a means of predicting who is likely to perform poorly at the end of the year without receiving additional assistance. However, we did include long-range prediction studies in Table 3 regardless of when the screener was administered (fall, winter, or spring) to include studies using recent advances in predictive validity methods.

Measures of Critical Aspects of Number Proficiency

Most research on screening measures in early primary grades (e.g., Lembke & Foegen, 2009; Methe, Hintze, & Floyd, 2008) has focused on discrete proficiencies, rather than deficiencies—an approach that seems more appropriate for universal screening measures. Although these measures are not designed to be comprehensive, when done well, their results may be related to students' performance on other critical aspects of mathematics. For example, a good measure of magnitude comparison may serve as an indicator of likely performance in place value or mental calculation. Most of these measures are easy to administer and typically take from as little as 1 min to 5 min to complete. Such measures could be used to quickly identify students whose mathematics achievement is either on track or at risk in one or more critical areas related to development of number sense/number proficiency, the most critical component of the early elementary grade mathematics curriculum (National Mathematics Advisory Panel, 2008).

However, as with any screening measure, these brief measures cannot provide a full diagnostic profile. As shown in Table 1, the predictive validity correlations are typically reasonable, but not high, and often not as high as comparable

TABLE 1

Predictive Validity of Screening Measures for the Primary Grades

Study	*Screening Measure*[a]	*Grade*	N[b]	*Outcome Measure*	*Predictive Validity*[c] (r)
	Magnitude Comparison				
Baglici, Codding, & Tryon (2010)	Test of Early Numeracy (TEN).	K	61	Timed Mathematics Computation	.02 (ns)
Chard, Clarke, Baker, Otterstedt, Braun, & Katz (2005)	Name the larger of two items: number sets 0 to 20.	K	436	Number Knowledge Test	.50
		I	483		.53
Clarke, Baker, Smolkowski, & Chard (2008)	Name larger of two items: number sets 0 to 10.	K	254	Stanford Early School Achievement Test	.62
Clarke, Gersten, Dimino, & Rolfhus (2012)	Name the larger of two items: number sets 0 to 20 for kindergarten and 0 to 99 for first grade.	K	323	Terra Nova	.49
		I	348		.62
Clarke & Shinn (2004)	Name larger of two items: number sets 0 to 20.	1	52	Woodcock Johnson Applied Problems	.79
				Timed Computation	.70
Lembke & Foegen (2009)	Name larger of two items: number sets 0 to 10 & 0 to 20 (i.e., 13:8).	K	44	Test of Early Mathematics Achievement-3	.35
		1	28		.43
Seethaler & Fuchs (2010)	Name the larger of two items: number sets 0 to 10.	K	196	Early Math Diagnostic Assessment:	
				Math Reasoning	.53
				Numerical Operations	.75
				Key Math-Revised:	
				Numeration	.34
				Estimation	.65

continues

TABLE 1 *Continued*

Study	*Screening Measure*[a]	*Grade*	N[b]	*Outcome Measure*	*Predictive Validity*[c] (r)
	Strategic Counting				
Baglici, Codding, & Tryon (2010)	Name the missing number in a string of numbers between 0 and 20.	K	61	Timed Mathematics Computation	.47
Clarke, Baker, Smolkowski, & Chard (2008)	Name the missing numeral from a string of numerals between 0 and 10.	K	254	Stanford Early School Achievement Test	.64
Clarke, Gersten, Dimino, & Rolfhus (2012)	Name the missing number in a sequence of numbers between 0 and 20 for kindergarten and 0 and 99 for first grade.	K	323	Terra Nova	.48
		1	348		.55
Geary, Bailey, & Hoard (2009)	Number Sets Test: child determines as quickly and accurately as possible if pairs or trios of object sets, Arabic numerals, or a combination of these matched a target number (5 and 9).	K	228	Wechsler Individual Achievement Test-II Numerical Operations subtest	.58
Lembke & Foegen (2009)	Name missing numbers in a pattern: Counting by 1s to 20, 5s to 50, and by 10s to 100 (i.e., 6 _ 8 9). Exact same items for kindergarten and first grade.	K	44	Test of Early Mathematics Achievement-3	.37
		1	28		.68
Methe, Hintze & Floyd (2008)	Students "count on" four numbers from a given number between 1 and 20 (e.g., experimenter says "8" and student says "9, 10, 11, 12")	K	64	Test of Early Mathematics Achievement-3	.46
	Word Problems				
Locuniak & Jordan (2008)	8-item story problems with 4 addition and 4 subtraction story problems.	K	198	Wechsler Intelligence Scale for Children-IV:	
				Digit Span Forward	.41
				Digit Span Backward	.31

continues

TABLE 1 *Continued*

Study	*Screening Measure*[a]	*Grade*	N[b]	*Outcome Measure*	*Predictive Validity*[c] (r)
		Fact Retrieval			
Bryant, Bryant, Gersten, Scammacca, & Chavez (2008)	TEMI: Addition/Subtraction (sums or minuends range from 0–18)	1	126	Stanford Achievement Test-10	.55
Clarke, Gersten, Dimino, & Rolfhus (2012)	ASPENS (working title): Basic Facts: Students are presented 40 problems that can be composed and decomposed in base-10 system	1	329	Terra Nova	.50

Note. All coefficients $p < .05$ unless noted otherwise; Screening measures: ASPENS = Assessing Student Proficiency in Early Number Sense, TEMI = Texas Early Mathematics Inventory. K = kindergarten.

[a] All measures were timed except Story Problems. [b] All study samples were from a single district except for Lembke & Foegen (2009) that sampled three districts in two states, and Clarke, Gersten, Dimino, & Rolfhus (2012) that sampled four districts in two states. [c] All predictive validity measured screeners administered in the fall and mathematics outcomes administered in the spring of that same year except Locuniak & Jordan (2008) which correlated the fall of kindergarten screening measure with criterion measures administered in the winter of first grade. Although the Seethaler & Fuchs (2010) calculated two predictive validity coefficients, only the fall and spring of kindergarten were used in this table.

TABLE 2

Multiple Number Proficiency Tests

Study	*Screening Measure*[a]	*Grade Screen*	n	*Outcome Measure*	*Grade Outcome*	*Predictive Validity*[b] (r)
Baker, Gersten, Flojo, Katz, Chard, & Clarke (2002) [c]	Number Knowledge Test: (takes about 10-15 min).	End of K	64	Stanford Achievement Test-9	1	.73
Clarke, Nese, Alonzo, Smith, Tindal, Kame'enui, & Baker (2011)	easyCBM: 45 items standardized computer-administered measure with items aligned to the National Council of Teachers in Mathematics Focal Point standards in mathematics.	1	145	TerraNova 3	1	.58
Jordan, Glutting, & Ramineni (2008)	Number Sense Brief: 33 items assessing counting, one-to-one correspondence, number recognition, nonverbal addition and subtraction.	K	204	Woodcock Johnson-III (Written Calculations and Problem Solving Subtest)	3	.63
Seethaler & Fuchs (2010)	Computation Fluency: 5-min timed assessment with 25 items of counting, addition and subtraction.	K	196	Early Math Diagnostic Assessment:	K	
				Math Reasoning		.57
				Numerical Operations		.62
				Key Math-Revised:		
				Numeration		.44
				Estimation		.68
Seethaler & Fuchs (2010)	Number Sense: 30 items.	K	196	Early Math Diagnostic Assessment:		
				Math Reasoning		.56
				Numerical Operations		.62
				Key Math-Revised:		
				Numeration		.40
				Estimation		.74

Note. K = kindergarten.

[a] The Number Sense Brief, the Number Knowledge Test, and the easyCBM measures were all un-timed assessments. [b] Although the Seethaler & Fuchs (2010) calculated two predictive validity coefficients, only the fall and spring of kindergarten were used in this table. [c] Baker et al. (2002) report predictive validity from spring of kindergarten to winter of first grade. Although this timeframe does not fall within the criteria we set, we have included the study because like the other studies it measures predictive validity over one year.

All coefficients $p < .05$.

TABLE 3

Diagnostic Utility Statistics and Receiver Operating Characteristics (ROC)

Study	*Screening Measure*	*Grade Screened*	n	*Outcome Measure*	*MLD or At Risk*	*Grade of Outcome*	*Sensitivity*	*Specificity*	*Area Under Curve*	*Duration of the Prediction*
Clarke, Nese, Alonzo, Smith, Tindal, Kame'enui, & Baker (2011)[a]	easyCBM: Standardized computer-administered measure with 45 items aligned to the National Council of Teachers in Mathematics Focal Point standards in mathematics.	1	145	< 25th percentile on the TerraNova 3	At risk Traditional	End of first grade	.83	.74	.83	1 year
				< 40th percentile on the TerraNova 3	At risk Liberal		.73	.73	.78	
Fuchs, Fuchs, Compton, Bryant, Hamlett, & Seethaler (2007)	Number Identification/Counting (Fuchs & Hamlett, 2005).	1	170	< 10th percentile on the WRAT 3-Arithmetic	MLD Calculation	End of second grade	.69	.79	.85	2 years
				< 10th percentile on Story Problems	MLD Word Problems		.70	.75	.81	
Geary, Bailey, & Hoard (2009)[b]	Number Sets Test: Child determines as quickly and accurately as possible if pairs or trios of object sets, Arabic numerals, or a combination of these matched a target number (5 and 9).	K	228	< 15th percentile on both the second and third grade Wechsler Individual Achievement Test-II Numerical Operations subtest	MLD	End of third grade	.66	.88	—	4 years

continues

TABLE 3 *Continued*

Study	*Screening Measure*	*Grade Screened*	n	*Outcome Measure*	*MLD or At Risk*	*Grade of Outcome*	*Sensitivity*	*Specificity*	*Area Under Curve*	*Duration of the Prediction*
Jordan, Glutting, Ramineni, & Watkins (2010)	Number Sense Brief: 33 items assessing counting, one-to-one correspondence, number recognition, nonverbal addition and subtraction.	K	204	Not meeting standards on the Delaware Student Testing in Mathematics (mix of concepts, procedures, and problem solving)	Low Achieving	End of third grade	.73	.85	.80	4 years
Seethaler & Fuchs (2010)	Quantity Discrimination	K	196	< 16th percentile on the EDMA Math Reasoning[c]	MLD Conceptual	End of first grade	.90	.66	.86	2 years
	Computation Fluency	K	196				.87	57.1	.80	
	Number Sense	K	196				.90	64.1	.84	
				< 16th percentile on the EDMA Numerical Operations subtest[d]	MLD Procedural		.89	.32	.69	
							.91	.35	.67	
							.88	.32	.69	

Note. K = kindergarten. QD = Quantity Discrimination. CF = Computation Fluency.

[a]Cut score following criteria proposed by Silberglitt and Hintze (2005). [b]True negatives = 109; False negatives = 22; True positives = 23; False positives = 4. [c]For the MLD-Conceptual: True Negatives = 103 on QD, 89 on CF, and 100 on Number Sense; False Negatives = 4 on QD, 5 on CF, and 4 on Number Sense; True Positives = 36 on QD, 35 on CF, and 36 on Number Sense; False Positives = 53 on QD, 67 on CF, and 56 on Number Sense. [d]For the MLD-Procedural: True Negatives = 44 on QD, 49 on CF, and 44 on Number Sense; False Negatives = 6 on QD, 5 on CF, and 7 on Number Sense; True Positives = 53 on QD, 54 on CF, and 52 on Number Sense; False Positives = 93 on QD, 88 on CF, and 93 on Number Sense.

measures in elementary school reading (Foegen, Jiban, & Deno, 2007). In addition, these measures are individually administered (at least at the current point in time). Although this mode may be preferable in kindergarten and the beginning of first grade, it is far more burdensome than computer-administered assessments or even pencil and paper assessments.

Four components of number sense/number competence deemed most important by cognitive psychologists include: (a) magnitude comparison (Booth & Siegler, 2006), (b) strategic counting (Geary, 2004), (c) the ability to solve simple word problems (Jordan et al., 2009), and (d) retrieval of basic arithmetic facts (Jordan, Hanich, & Kaplan, 2003). Table 1 provides descriptions of key elements of the literature base among the four components of number sense/number competence.

Magnitude Comparison. Magnitude comparison is the ability to discern which number is the greatest in a set, and to be able to weigh relative differences in magnitude efficiently (e.g., to know that 11 is a bit bigger than 9, but 18 is a lot bigger than 9). As children develop a more sophisticated understanding of number and quantity, they are able to make increasingly complex judgments about magnitude. Riley, Greeno, and Heller (1983) found that, given a hypothetical scenario with a picture of five birds and one worm, most preschoolers could answer questions such as, “Suppose the birds all race over and each one tries to get a worm. Will every bird get a worm?” Their answers demonstrate a gross magnitude judgment that there are more birds than worms. But given a specific question about magnitude, for example, “How many birds won’t get a worm?” (p. 169), most preschoolers could not answer correctly. The ability to make these finite types of magnitude comparisons is a critical underpinning of the ability to calculate, and, in the view of Okamoto and Case (1996) as well as Booth and Siegler (2006), represents the evolution of an increasingly sophisticated and accurate mental number line, as discussed above.

A number of research teams have designed and tested similar measures of magnitude comparison for kindergarten and first grade (See Table 1). All measures included a timed element, but the range of numbers used in the materials varied in response to potential concerns about floor or ceiling effects. One of the first efforts to develop a measure of magnitude comparison was by Clarke and Shinn (2004), who tested a timed magnitude comparison measure with first-grade students using fall screening to predict performance on the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R) Applied Problems subtest (Woodcock & Johnson, 1989). Predictive validity was .79, which is quite high. Clarke, Baker, Smolkowski, and Chard (2008) extended the work to a kindergarten sample, only including numbers between 1 and 10, rather than 1 and 20. Predictive validity was .62 with a standardized achievement test.

Table 1 presents additional data on predictive validity of magnitude comparison measures. Coefficients were fairly consistent, with a median of .62 for first grade and .50 for kindergarten. Many studies suffer from the limitation of using only one site, with the exception of Clarke, Gersten, Dimino, and Rolfhus (2012) and Lembke and Foegen (2009). However, the fact that kindergarten and first-grade findings were replicated across multiple studies in multiple sites does indicate great promise for a timed measure of magnitude comparison.

Strategic Counting. The ability to understand how to count efficiently and use counting strategies is fundamental to developing mathematical understanding and proficiency (Siegler & Robinson, 1982). Geary (2004) notes that weak ability in counting strategies is a key indicator of which young students are likely to have difficulty learning mathematics. In most cases, competence in counting strategies is strongly related to burgeoning knowledge of number properties. Once a child possesses the “count on” strategy, if asked “what is 9 more than 2?” she will automatically know that it is much more efficient to reverse the problem to 2 more than 9, and simply “count on” from 9. Counting on from the larger addend is important for learning addition and subtraction number combinations, and grasping the count on strategy demonstrates the beginnings of a grasp of the commutative property of addition.

A number of researchers have developed strategic counting measures that require students to identify the missing number from a sequence of numbers. All measures included a timed element, but different ranges of numbers were used. The Clarke et al. (2008) Missing Number kindergarten measure used numbers between 1 and 10, and their first-grade measure used numbers between 1 and 20. Lembke and Foegen (2009) developed a measure of strategic counting in which students were given 1 min to identify a missing number from a sequence of four consecutive numbers. They also included items (20% of the items) that assessed strategic counting by 5s and 10s (e.g., 5 10 __ 20). The range of numbers used was up to 20 for count by 1s, up to 50 for count by 5s, and up to 100 for count by 10s. Unlike other researchers, they used the same items for kindergarten and first grade. The predictive validity was weak for kindergarten (.37), in fact the weakest in the set of studies. In contrast, it was moderately strong for first grade (.68), suggesting that kindergartners and first graders require different sets of items in screening measures.

Each research team found moderate concurrent and predictive validities (range = .37–.68) and strong reliabilities (range = .59–.98). Specifically, the predictive validities are quite high for first grade with a median coefficient of .62, but in the low to moderate range for kindergarten with a median of .475. This measure thus does not seem a suitable screener for the beginning of kindergarten.

Word Problems Involving Simple Arithmetic Operations. Jordan, Levine, and Huttenlocher (1994) found that although adults often think that children have a hard time solving word problems, young children, in fact, find them easier than even simple number sentences. In other words, before formal schooling, children can much more easily tell you how many sheep are left if you start out with 9 and lose 2 than they can tell you that 9 minus 2 is 7.

For that reason, simple word problems have been added to early screening batteries in recent years (Fuchs et al., 2007; Locuniak & Jordan, 2008). Locuniak and Jordan created a simple eight-item story problem measure with four addition and four subtraction story problems. Performance on the story problem measure in the fall of kindergarten was moderately related to performance on a measure of calculation fluency at the end of second grade (.51), quite high for predicting performance over the course of 3 school years. Performance was less related between the word problem measure and the digital span forward and backward on the Wechsler Intelligence Scale for Children-IV (WISC-IV; Wechsler, 2003).

Retrieval of Basic Arithmetic Facts. Some of the earliest research on mathematics difficulties focused on correlates of students in the upper elementary grades who were identified as demonstrating a learning disability by school personnel. One consistent finding (Goldman, Pellegrino, & Mertz, 1988) was that students who struggled with mathematics in the elementary grades were unable to automatically retrieve addition and subtraction number combinations. Research seems to indicate that although students with learning disabilities in mathematics often make good strides in terms of facility with algorithms, procedures, and simple word problems, severe deficits remain in their retrieval of basic combinations (Geary, 2004; Jordan et al., 2003). These deficiencies suggest underlying problems with what Geary calls semantic memory (i.e., the ability to store and retrieve abstract information efficiently). This ability appears to be critical for students to succeed in mathematics and, ultimately, to understand mathematics. Jordan et al. (2003), however, argue that poor fact retrieval has its roots in weak number sense. It is difficult for children to become automatic with addition and subtraction number combinations when they do not have a good sense of relations between and among numbers and operations. The ability to solve number combinations involving addition and subtraction, even at the beginning of kindergarten, is considered a powerful predictive measure of mathematics achievement through third grade (Jordan et al., 2009).

It is difficult for children to become automatic with addition and subtraction number combinations when they do not have a good sense of relations between and among numbers and operations.

Measures of fact retrieval appear to be promising based on one study by Bryant et al. (2008), who designed an addition and subtraction fact screener for first- and second-grade students. Students had 1 min to complete basic facts problems. Concurrent validity correlations with the Stanford Achievement Test-Tenth Edition (SAT-10; Pearson, 2003) were .55 for first-grade students and .59 for second-grade students. These data suggest that a fact retrieval measure would be a sensible addition to a screening battery in the second grade and possibly first grade. In fact, recent research (Clarke et al., 2012) examined the predictive validity of a timed fact retrieval measure with the widely used mathematics achievement test Terra Nova (CTB/McGraw-Hill, 2008) for first-grade students and found a correlation coefficient of .50.

Each of the four competencies discussed previously appears reasonable for use in early screening. Each of the measures discussed is brief and easy to administer, important characteristics of a screening measure.

MULTIPLE NUMBER PROFICIENCY MEASURES

Another promising approach is the use of measures that cover multiple, but related number competencies that young children need in order to be successful in mathematics. Much of the research in this area is quite new, but appears equally promising as the single proficiency measures. Research to date demonstrates that measures encompassing multiple aspects of number competence, such as the Number Knowledge Test (NKT; Okamoto & Case, 1996) and Number Sense Brief (Jordan, Glutting, & Ramineni, 2008), tend to demonstrate somewhat stronger predictive validity than the briefer, single proficiency measures. Table 2 provides a description of multiple number proficiency test studies and their reported predictive validity.

The NKT is an individually administered measure that is one of the earliest attempts to assess students' procedural and conceptual knowledge related to whole numbers. The NKT includes a number of the critical proficiencies described previously, such as the ability to make magnitude comparisons, count, and use basic arithmetic operations in multiple formats including word problems that are read to the student. The NKT takes about 10 to 15 min to administer and consists of four levels of increasing difficulty. For example, children at the second level compare the numbers 5 and 4 and identify the bigger number. The same problem type is presented at the third level using the numbers 19 and 21.

Baker et al. (2002) explored the ability of the NKT administered at the end of kindergarten to predict mathematics achievement on the Stanford Achievement Test-Ninth Edition (SAT-9; Pearson, 1996) at the end of first grade. The predictive validity coefficient of the NKT was .73 for Total Mathematics. Note that this lengthier multiple-proficiency measure tends to demonstrate slightly higher predictive validity than many of the briefer measures discussed earlier.

Finally, item response theory (IRT) was used to establish the internal consistency and reliability of the NKT and to examine the extent to which the four levels of the NKT fit item difficulties. The IRT reliability was .93. Descriptive analyses revealed that most of the items fit the levels established by Okamoto and Case (1996), although a few were misplaced and a paucity of items were at the easy level of difficulty, indicating that the measure would not necessarily be sensitive to growth for students at the lower end of the distribution.

More recently, Jordan et al. (2008) developed a screening battery based on the same theoretical and empirical underpinnings of the Locuniak-Jordan research but much more brief and efficient, with an administration time of approximately 15 min. Test-retest reliability ranged from .61 to .86 with a predictive validity of .63 from kindergarten administration to student mathematics achievement, measured by the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) in third grade.

Seethaler and Fuchs (2010) administered a single proficiency measure, a magnitude comparison (Chard et al., 2005), and a multiple proficiency measure (the Number Sense, created by the authors) in September and May of kindergarten. At the end of first grade, conceptual and procedural outcomes were measured on The Early Math Diagnostic Assessment (EMDA; The Psychological Corporation, 2002) and the KeyMath-Revised (KM-R; Connolly, 1998). Comparisons of single and multiple proficiency screeners, fall versus spring kindergarten screening, and conceptual versus procedural outcomes were conducted using logistic regression and ROC analyses. Results indicated that single and multiple proficiency screeners produced good and similar classification accuracy at the fall and spring screening occasions on the conceptual outcome. Interestingly, each of their screeners classified future conceptual math difficulties (MD) status with significantly greater accuracy than future procedural MD status. During the fall of kindergarten, the area under the curve (AUC) for the three screeners ranged from .80 to .86 for the EMDA Math Reasoning subtest, indicating good predictive utility for conceptual MD status, whereas AUC for the EMDA Numerical Operations subtest ranged from .67 to .69, indicating poor predictive utility for procedural MD status. One possibility is that strategic counting would theoretically seem to be a strong predictor of computational proficiency. Another possible explanation, proposed by the authors, is that numerical operations are not stressed in kindergarten and first grade, whereas foundational mathematical concepts are heavily stressed.

SCREENERS ALIGNED TO CURRICULUM STANDARDS

A different approach to screening has been developed by Fuchs and colleagues (e.g., Fuchs, Fuchs, & Zumeta, 2008) and Clarke and colleagues (Clarke et al., 2012; Clarke et al., 2011). Typically, this screening approach consists of a group administered paper and pencil test containing items that represent the current year's curricula scope and sequence (derived, for example, from current state standards or National Council of Teachers of Mathematics [NCTM] Focal Points). A strength of these measures is that they can be quickly administered and scored by machine. They also possess strong face validity because of their focus on key curriculum topics for the current year. Typically, these measures demonstrate acceptable test-retest, inter-rater, and alternate form reliability above .80. The concurrent and predictive validities of these measures are between .50 and .60 (See Foegen, et al. [2007] for an extensive review). There are some instances of their use in first grade (e.g., Seethaler & Fuchs, 2010 and Clarke et al., 2011).

OTHER MEASURES TO CONSIDER INCLUDING IN A MORE COMPREHENSIVE SCREENING BATTERY

One might think that only mathematics measures should be used in screening for potential mathematics difficulties. However, recent research on early identification of students with problems learning mathematics has discovered that working memory and student engagement may also be useful in predicting problems in mathematics. For that reason, we discuss both in this article.

Working Memory

Recent syntheses of the literature on mathematics disabilities (e.g., Desoete, Ceulemans, Roeyers, & Huylebroeck, 2009; Geary, 2004) observe that, in addition to problems with magnitude comparison, counting strategies, and computational strategies, these students often display deficits in working memory (Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007; Swanson & Beebe-Frankenberger, 2004) and problems with visual-spatial memory and elaboration (e.g., Geary, 2004). Contemporary research consistently demonstrates the importance of working memory (Baker et al., 2002; Locuniak & Jordan, 2008; Swanson & Beebe-Frankenberger, 2004) in understanding mathematical proficiency at many different age levels. Working memory is often measured by a reverse digit span task, that is, a task requiring a student to repeat a set of numbers read to him (e.g., 9, 4, 17, 8) in precisely the reverse order (i.e., 8, 17, 4, 9).

Working memory seems to function less efficiently as a screening measure than as a variable that adds precision to a set of other predictors (e.g., Baker et al., 2002; Locuniak & Jordan, 2008). The relationship between working memory and number sense appears to be complex. We believe that working memory relates to mathematics proficiency because in mathematics, students are asked not only to remember, but also mentally "juggle" several bits of abstract information (e.g., basic facts, positions of numbers on a mental line, computational procedures, etc.). Students with weak number sense need to rely more on working memory. Students weak in both areas are likely to struggle and future research should target optimal intervention strategies for this group of students.

Student Engagement and Attentiveness

The relationship between teacher ratings of a student's attentiveness and at-risk status in mathematics also appears to be consistent (Bodovski & Farkas, 2007; DiPerna, Lei, & Reid, 2007; Fuchs et al., 2007). Recent research on early identification of students with difficulties learning mathematics suggests that measures of a student's attentiveness during academic instruction is a solid predictor of future mathematics achievement. Utilizing the Early Childhood Longitudinal Study (ECLS–K) database, a number of researchers have begun to explore the relationship between student engagement and achievement growth in mathematics.

Bodovski and Farkas (2007) examined the growth rates from the fall of kindergarten to the spring of third grade with a nationally representative sample of 13,043 students. ECLS–K measured student engagement by asking teachers to complete a six-item survey rating each student's attention, persistence with tasks, and demonstration of learning independence. As expected, higher achieving students displayed greater levels of engagement. A secondary analysis of the lowest quarter of students found that within this group, students with the greatest need showed lower rates of achievement growth and engagement over time. Across all grade levels tested (i.e., K–5), the attention measure contributed a unique proportion of the variance in outcomes, beyond initial skill status. This effect was striking because the impact of student engagement was greater than time spent on instruction and the effect showed the greatest impact for the lowest achieving students. This finding suggests that interventions for students with problems in mathematics might seriously consider adding a component that promotes attentiveness to academic tasks and activities.

DiPerna et al. (2007) also examined the relationship between student engagement and achievement growth in mathematics using the ECLS-K database from kindergarten entry to third-grade exit with a nationally representative sample of 3,240 students. Teachers' ratings of students' academic attentiveness at the beginning of kindergarten played a role in predicting subsequent mathematics achievement, with predictive validity correlation coefficients ranging from .28 to .35. Fuchs et al. (2007) also found that teachers' appraisal of attentiveness was a significant predictor of future mathematics achievement.

Although these correlations are not nearly as large as those of the mathematics screening measures, they suggest that ratings of attentiveness could be added to a screening battery to create some type of composite score. They also suggest that many students requiring some type of early intervention in mathematics might also struggle with maintaining attention to academic tasks for sustained amounts of time.

As researchers build more sophisticated models for early identification using more extensive batteries (e.g., Fuchs et al., 2007; Locuniak & Jordan, 2008), we may begin to generate more precise screening methods. The more sophisticated models may take into account measures of working memory, attentiveness, and perhaps other cognitive variables.

PREDICTIVE VALIDITY AND DIAGNOSTIC CLASSIFICATION ACCURACY

Predictive Validity

The main method for evaluating the measures discussed in this article was to examine predictive validity using Pearson correlations over the course of a school year. (A small number of studies were conducted over a longer timeframe, e.g., 2 to 3 years. We discuss these separately.) All of these studies involved multiple proficiency measures or studies that used diagnostic utility statistics. When designing screening measures in mathematics, a critical variable to consider is the extent to which performance on those measures relates to later performance in mathematics. For example, a student's score on a first-grade screening measure would need to accurately predict difficulty in mathematics at the end of first grade, and ideally performance a year later as well. Assessments that show evidence of predictive validity can aid in instructional decision making. If evidence indicates that a score below a certain threshold on a kindergarten or beginning of first-grade measure of mathematics predicts later problems, then schools and teachers can use that information to allocate resources for instructional or intervention services to those students.

This approach provides reasonable estimates of how well the distribution of scores on the screener corresponds with the distribution of students' performance on a lengthier achievement test in mathematics administered later in the school year. However, for screening measures, we are more interested primarily in one group of students—the 15% to perhaps 30% who are at risk. Ultimately, a screening measure rises or falls based on how well it is able to pinpoint which students need additional help. For that reason, recent screening research has begun to use measures of classification accuracy (e.g., Clarke et al., 2011; Jordan, Glutting, Ramineni, & Watkins, 2010; Mazzocco & Thompson, 2005; Seethaler & Fuchs, 2010). Because this information is relatively new to special education research (this is less true for psychology research), and because the concepts—classification accuracy, specificity, sensitivity, and ROC curves—are relatively new and often not well understood, we explain them below.

Classification Accuracy

Classification accuracy refers to the degree to which the screener provides correct classifications of children who require additional assistance in mathematics. There are two types of mistakes a screener or screening battery can make. The first is to miss students who truly need help, that is, to create false negatives. To assess this risk, we report on the screener's sensitivity. The second type of classification mistake is to falsely identify students as needing help when in fact they do not require additional instruction or assistance. This group of students is called *false positives.* To assess this risk, we report on the screener's specificity.

Earlier, the field focused heavily on high sensitivity, and that remains a major concern in much of the published research (e.g., Seethaler & Fuchs, 2010). However, the response to intervention (RTI) research community has become increasingly aware of the phenomenal waste of resources that comes with false positives, and there has been more focus on specificity (see, for example, Silberglitt & Hintze, 2005).

Sensitivity: Ensuring That No One "Falls Through the Cracks"

Jordan provides a practical definition for the term sensitivity: "Sensitivity is the proportion of individuals with a disorder (e.g., individuals with low achievement or learning disabilities) who are correctly identified by a positive test finding" (Jordan Glutting, Ramineni, & Watkins, 2010, p. 184). Researchers have used various working definitions of how to determine what it means to "require additional assistance" in an academic area, realizing that any of these operational definitions are, at best, educated guesses.

There is no common standard for determining what the term "at risk" or "would benefit from intervention" means. The problem is hardly unique to education. Public health officials grope with this issue as they determine categories such as "at risk for heart disease" or "at risk for a stroke," and change criteria to reflect evolving definitions of at risk.

Some researchers have focused less on RTI and more on valid early identification of students who possess a disability in mathematics. These researchers often use a criterion of the 10th percentile (e.g., Fuchs et al., 2007; Morgan et al., 2009), reasoning that about one student in ten has a learning disability in mathematics or is at strong risk for developing a disability in mathematics in the future. Others focus on using a screening measure or battery to determine which students might benefit from an intervention. These very different criteria for classifying a student as "at risk" have a profound impact on any discussion of sensitivity, and the professional literature often does not make this distinction clear.

During the initial development of screening measures in mathematics and reading, researchers often argued that it was most important to "catch kids early," and they feared letting any at-risk students through the screening battery. Thus, they often set a high criterion, such as the 40th percentile or below, to avoid missing any students who might have a true problem. However, there are costs to this approach. The formal means for assessing those costs is specificity.

Specificity and the Concern for Wasted Resources

Often neglected but equally important is knowledge of whether students recommended for Tier 2 intervention would succeed without being provided with any particular intervention. We refer to this as the *specificity* of the screener. In the zeal to make sure that we do all we can to provide early intervention to students in reading and mathematics, this topic has been neglected until recently (Gersten et al., 2009; Silberglitt & Hintze, 2005). At this time, there is no consensus or commonly used convention for reporting what it means to "succeed without any additional intervention." Deciding what performance criterion will be used is a key issue that is not easy to resolve.

Weak specificity indicates that a screening measure or battery is over-identifying students and thus providing services to students who do not need them. Students who are misclassified as needing help even though they don't are called *false positives.* False positives are problematic because resources may be wasted (Gersten et al., 2009) by providing extra intensive intervention to students who do not need such help, as often happens in the field of reading (e.g., Jenkins & O'Connor, 2002). For schools with finite resources, this is particularly vexing. Resources spent on providing interventions to students who do not need them are thus not available to be spent on other valuable services. False positives are not only taxing on schools but can also be detrimental to parents and students " . . . given the generally 'chaotic' nature of early achievement and the increased possibility of falsely identifying students as being 'at-risk' when they are merely distracted, anxious, or unfamiliar with the testing protocols" (Bryant et al., 2011, p. 9). Students and parents may suffer adverse effects of thinking the child has a disability when in fact the student has no disability whatsoever.

Can We Balance These Two Concerns? ROC Curves as a Potential Tool

A measure with perfect sensitivity ensures that all students who require intervention receive extra support. A measure with perfect specificity ensures that schools do not spend resources on students who do not need extra support. However, measurement in education, medicine, psychology, and most human endeavors is far from perfect and consists of a series of compromises and balances. For screening, we need to balance two goals – accurately detecting which students require early intervention (sensitivity) and detecting only those students who require additional help (specificity).

Here we face a bit of a paradox. The more we increase sensitivity, the more we try to ensure that we do not miss any students who might need intervention, but in doing so the more we decrease specificity. Thus, development and refinement of effective screening measures requires a delicate balance. The selection of a cut score (the number at which a score at or above classifies the student as not at risk and a score below classifies the student as at risk) affects both sensitivity and specificity in a reciprocal fashion (e.g., setting the cut score to have higher sensitivity leads to lower specificity).

In the past, most researchers simply skirted the issue and reported Pearson correlations of predictive validity. As a field, we are only beginning to develop conventions for reporting on sensitivity and specificity. One widely used tool in evaluating the utility of a diagnostic instrument is the ROC curve (see Table 3 for a summary of studies utilizing ROC analyses). Although relatively new to the area of mathematical screening, the classification accuracy of diagnostic tests has long been of interest in medicine. "By systematically using all possible cut scores of a test and plotting the true-positive rate (i.e., sensitivity) against the false-positive rate (i.e., 1-specificity) for each cut score, diagnostic validity can be displayed for the full range of the test's scores" (Jordan, Glutting, Ramineni, & Watkins, 2010, p. 184). In essence, the ROC curve plots sensitivity vs. 1-specificity and illustrates the inverse relationship between sensitivity and specificity; changing the cut score to improve one will lower the other.

Example of an ROC Analysis

If a researcher chose a score on a screener that would correctly identify every child who scored below a given criterion on subsequent mathematics achievement tests (e.g., the 10th percentile for MLD or the 25th for at-risk status), they would invariably include many false positives, thus resulting in unacceptably low specificity. ROC analyses can help inform these decisions.

ROC analyses select the cut score mathematically to maximize sensitivity and specificity in a balanced way. In an ROC analysis, the outcome is specified a priori. For example, Fuchs et al. (2007) specified a score below the 10th percentile on various mathematics achievement measures as representing MLD. Then, using the AUC in the ROC analysis, which provides a metric for accuracy of group discrimination based on the a priori cut score, they were able to identify how accurately the number sense measure discriminated between the MLD and non-MLD groups based on their performance. By examining the ROC curve, one can actually examine the impact of various cut scores on accurate group identification (e.g., MLD or at risk for MLD). However, although an ROC analysis increases sensitivity, it does not concomitantly increase specificity. The role of the cut score is an integral one. Recent research demonstrates how ROC analyses can assist researchers in development of accurate measures, but there are trade-offs that necessitate acknowledgement.

Contemporary Research Using ROC Analyses: The Search for an Ideal Balance

Earlier studies that pioneered the use of ROC analyses typically reported the AUC and noted whether it was higher than .80. If so, they reported that the AUC was "good" following a standard of clinical significance, with .80 to .89 being "good" and .90 to 1.00 being "excellent" (Cicchetti, 2001). In contrast, Seethaler and Fuchs (2010) perform a much more sophisticated, useful analysis. In addition to reporting sensitivity, specificity, and the AUC, the authors provide the number of students incorrectly identified by the screener as MD (false positives), the number of students incorrectly identified as non-MD (false negatives), the number of students correctly identified as MD (true positives), and the number of students correctly identified as non-MD (true negatives). They did this analysis separately for students classified as MD-conceptual and MD-procedural. Students received an MD-conceptual designation if they scored below the 16th percentile on the EMDA Math Reasoning subtest and MD-procedural if they scored below the 16th percentile on the EMDA Numerical Operations subtest. As can be seen in Table 3, Seethaler and Fuchs report sensitivity of 89.8% for the quantity discrimination (Clarke & Shinn, 2004) screener in predicting MD-procedural status. However, the number of students incorrectly identified as MD was also high (93). In other words, despite high values of sensitivity, nearly half of the students testing positive were actually false positives, resulting in a specificity of 32.1%. Thus, simply relying on sensitivity can produce misleading conclusions about the diagnostic accuracy of a test.

Clarke et al. (2011) performed a similar analysis on the first-grade version of a newly designed measure called easyCBM (http://www.easycbm.com/). In this case, they used a criterion developed by Silberglitt and Hintze (2005). These researchers specify a specific means for use of an ROC curve. This entails (a) only including cut scores so that both specificity and sensitivity are equal or greater than .70, and (b) performing an intricate titration process so that sensitivity is increased as much as possible while specificity remains at least .70. This process is described in detail in Clarke et al. (2011) and demonstrates a promising method for balancing the need to correctly identify as many students who need help as possible while not casting such a wide net that students who would do fine without help are given costly assistance.

DISCUSSION

The recent report on early childhood mathematics learning (National Research Council, 2009) concluded:

> Further exploration is needed to better understand what early number competencies are predictive of future success in mathematics. Such research can help identify children at risk for learning difficulties or disabilities in mathematics . . . [and] develop targeted interventions for such children and test their effectiveness. (p. 350)

The authors were primarily addressing preschool, but the need is as critical for students in the primary grades. The longitudinal studies documenting the persistence of mathematics disabilities and difficulties in learning mathematics from kindergarten to the upper elementary grades create a compelling case for future research on development and refinement of valid universal screening measures for students in the primary grades.

In the remainder of this section, we highlight several areas for future research. We also note pragmatic issues faced by school personnel attempting to implement some of the screening measures discussed earlier.

Grappling With the Concept of Risk Status

Determining risk status is as much an art as a science. This is true in all fields. However, it remains a vexing issue in the area of mathematics for several reasons. A source of confusion in the field is that criteria for determining at-risk status in mathematics have varied from below the 25th percentile on a normed mathematics measure (Locuniak & Jordan, 2008) to below the 10th percentile (Fuchs et al., 2007; Morgan et al., 2009). In the first case, the researchers mirrored what a school district might do: cast a relatively broad net to ensure that all students who may need intervention receive it. However, as districts have learned from experiences with RTI in reading, and as researchers have begun to consistently note, there are real drawbacks to casting too broad a net. For one thing, a good deal of time and money is wasted because intervention is provided to students who would do fine without it. In addition, resources that are usually sorely needed are pulled away from intermediate grades and middle school.

Potential Limitations of ROC Analyses

Because there remains no clear definition of what constitutes “at risk,” decisions made regarding at-risk classification are often complex. Normally, researchers select one specific cut score and students who do not achieve that predetermined score are determined likely to be at risk in the area targeted by the assessment. However, researchers have a lot of discretion in selecting the precise score to use. The decisions have a significant impact on the classification accuracy of any screening instrument selected and the number of students identified for additional support in mathematics. The recent research focusing on classification accuracy using ROC analyses to evaluate the sensitivity and specificity of a screening system is certainly a step forward over simply presenting a predictive validity correlation coefficient. However, ROC analyses are based on various, usually unstated, mathematical assumptions (i.e., normality of noise [or error] distributions). A recent article by VanDerHeyden (2011) provides one of the most probing discussions of the limitations of sensitivity and specificity for both practitioners and researchers using these analyses. The author proposes a more probabilistic, Bayesian approach, inspired by the work of Robyn Dawes (1962). VanDerHeyden highlights the importance of considering positive predictive power, that is, the probability that a score below benchmark is an accurate indicator of risk. She also provides important cautions:

> Sensitivity and specificity offer little information about the value of a test finding for ruling-out or ruling-in a condition. Generally, predictive power quantifies the value of a test finding for ruling in (positive predictive power) and ruling-out (negative predictive power) a condition in a way that is easily interpreted and used. (VanDerHeyden, 2011, p. 342)

In other words, predictive power can tell a teacher or psychologist whether a particular student is really at risk for learning problems in math. However, positive and negative predictive power estimates are problematic because the field has no precise definition of what it means to be "at risk" for a learning problem in math or to possess a math learning disability.

Advances in the Use of ROC Analyses

Ultimately, we envision that ROC can help us select a benchmark, but we then need to ensure that our cut score demonstrates adequate validity in terms of consequential validity (Gersten, Keating, & Irvin, 1995; Messick, 1980), that is, use of a given screening procedure must be linked to increases in the mathematics performance of students at the lower end of the distribution in math. We need to study the impacts of implementing specific screening measures or batteries and specific benchmarks on practice, using occasional case study and descriptive research.

The concern for false positives (low specificity), especially in the early grades, has received increased attention in both reading and mathematics. Recently, Siegel, Fuchs, O'Connor and Vaughn (2011) experimented with a method for reducing this rate in the primary grades. Although the impact was not statistically significant, this does seem to be a promising method for future study. Students who score below the cut score in the fall are not immediately placed in a Tier 2 intervention. Rather, they continue to receive only typical classroom instruction, but their progress is monitored closely (e.g., weekly) for a period of 6 to 8 weeks. Only those with unacceptably low rates of progress receive Tier 2 interventions. In our view, this type of approach warrants further research and is worthy of serious consideration.

We are only beginning to understand how to use the concepts of sensitivity, specificity, and classification accuracy in our research and how these analyses can provide critical information for districts or schools making decisions about what type of screening measures to use. For example, a district may decide to select a more efficient screening measure over a more comprehensive (less efficient) measure if the shorter measure demonstrates similar rates of accurate classification. Earlier studies merely reported that AUC was over .80 and concluded it was good based on clinical significance criteria and stopped there. The contemporary research by Seethaler and Fuchs (2010) and Clarke et al. (2011) seems to be a timely advance in the increasingly sophisticated use of ROC to balance the competing demands of specificity and sensitivity—of wasting resources versus letting students fall through the cracks. Despite its limitations, ROC remains a useful tool for establishing criteria for cut scores and can now be used by individual school districts (or individual schools) if appropriate technical support is provided.

Our Perspective on the Future

With the advent of the Common Core State Standards (http://www.corestandards.org/the-standards/mathematics) and increased use of technology, notions of efficiency change. Whereas even 5 to 10 years ago, the main consideration for efficiency was how long a test took to administer and score, with the use of technology, scoring can be done almost automatically. Testing, at least beginning sometime in first grade, would not require high degrees of adult supervision. Most importantly, use of IRT allows technology to calibrate screening measures more precisely.

We are only beginning to understand how to use the concepts of sensitivity, specificity, and classification accuracy in our research and how these analyses can provide critical information for districts or schools making decisions about what type of screening measures to use.

At this point in time, we understand a good deal more about what comprises a comprehensive assessment battery, but are less certain of the elements of an efficient assessment battery. A crucial criterion for use of a screening measure is efficiency. Jordan et al. (2008) have developed an efficient 33-item untimed screening measure that has good predictive validity, and the NKT is a relatively efficient screening tool, typically taking 10 to 15 min to administer. Both of these are far more comprehensive than measures of one component of number sense, such as magnitude comparison. However, the realities of universal screening require use of the most efficient measures.

It is critical to note that the development of mathematical thinking is broader than the set of skills assessed by any one of the measures described previously, which focus almost exclusively on the domain of numbers. Some challenges the field will face will be to explore the role of student performance in other critical areas (e.g., geometry), to determine how to measure performance in those domains, and to determine their relationship to proficiency in algebra and other advanced mathematical topics (G. J. Duncan, personal communication, May 7, 2011). As our understanding of mathematical development advances, so too should our design of screening instruments that reflect the complexity of mathematics. However, the insights about the importance of the sophistication of an individual's mental number line as a sensitive snapshot of mathematical development made by Okamoto and Case (1996) remain robust, as the evidence described in this article demonstrates.

Each year in school brings about new challenges for students and new material to master in order to further their mathematical understanding and build a foundation for future content. Because the demands of the mathematics curriculum continue to change over the years, it is possible that certain students may initially learn math at acceptable levels only to experience problems once content moves to a more abstract level (e.g., with the introduction of decimals, improper fractions, ratios and proportions, negative numbers). Therefore, as in the reading field (Scarborough, 2001), we will likely see students whose performance in mathematics is acceptable in the primary grades, but deteriorates in later grades.

Future research needs to address several critical areas. The first is valid screening measures for Grades 3 and above, using IRT and important policy frameworks such as the Common Core Standards as a basis. Another potential research area is measurement of skills related to geometry and an examination of whether there are precursors of geometry proficiency that are different than those of proficiency with number concepts and operations.

Last, we would be remiss if we did not emphasize that the collection of screening data in and of itself does not change student outcomes. Any advances that schools make in screening students in mathematics must occur alongside efforts to improve instructional practices and to develop effective interventions. The body of research on this topic is sparse, but expanding rapidly.

Any advances that schools make in screening students in mathematics must occur alongside efforts to improve instructional practices and to develop effective interventions.

REFERENCES

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the research on the tests of early numeracy: Longitudinal analyses over two school years. *Assessment for Effective Intervention, 35*(2), 89–102. doi:10.1177/1534508409346053

Baker, S., Gersten, R., Flojo, J., Katz, R., Chard, D., & Clarke, B. (2002). *Preventing mathematics difficulties in young children: Focus on effective screening of early number sense delays* (Tech. Rep. No. 0305). Eugene, OR: Pacific Institutes for Research.

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

Bodovski, K., & Farkas, G. (2007). Do instructional practices contribute to inequality in achievement? The case of mathematics instruction in kindergarten. *Journal of Early Childhood Research, 5,* 301–322. doi:10.1177/1476718X07080476

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 41*, 189–201. doi:10.1037/0012-1649.41.6.189

Bryant, D. P., Bryant, B. R., Roberts, G., Vaughn, S., Pfannenstiel, K. H., Porterfield, J., & Gersten, R. (2011). Early numeracy intervention program for first-grade students with mathematics difficulties. *Exceptional Children, 78*, 7–23.

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. (2008). Mathematics intervention for first and second grade students with mathematics difficulties: The effects of tier 2 intervention delivered as booster lessons. *Remedial and Special Education, 29*(1), 20–32. doi:10.1177/0741932507309712

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*, 3–14. doi:10.1177/073724770503000202

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23*(5), 695–700. doi:10.1076/jcen.23.5.695.1249

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education, 29,* 46–57. doi:10.1177/0741932507039694

Clarke, B., Gersten, R., Dimino, J., & Rolfhus, E. (2012). *Assessing student proficiency in early number sense (ASPENS)* [Measurement instrument]. Longmont, CO: Cambium Learning Sopris.

Clarke, B., Nese, J. F. T., Alonzo, J., Smith, J. M., Tindal, G., Kame'enui, E., & Baker, S. (2011). Classification accuracy of easyCBM first grade mathematics measures: Findings and implications for the field. *Assessment for Effective Intervention, 36,* 243–255. doi:10.1177/1534508411414153

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

Connolly, A. J. (1998). *KeyMath-Revised.* Circle Pines, MN: American Guidance Service.

CTB/McGraw-Hill. (2008). *TerraNova Third Edition Complete Battery.* Monterey, CA: Author.

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology, 26,* 422–424. doi:10.1037/h0044612

Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review, 4*(1), 55–66. doi:10.1016/j.edurev.2008.11.003

DiPerna, J. C., Lei, P., & Reid, E. E. (2007). Kindergarten predictors of mathematical growth in the primary grades: An investigation using the early childhood longitudinal study—kindergarten cohort. *Educational Psychology, 99*, 369–379. doi:10.1037/0022-0663.99.2.369

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

easyCBM (2012). Curriculum-based measurement solutions for every tier [Assessment software]. Retrieved from http://www.easycbm.com

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41,* 121–139. doi:10.1177/00224669070410020101

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.

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). Response to intervention: A strategy for the prevention and identification of learning disabilities. In E. L. Grigorenko (Ed.), *Educating individuals with disabilities: IDEIA 2004 and beyond* (pp. 115–135). New York, NY: Springer.

Fuchs, L. S., & Hamlett, C. L. (2005). *Number identification/counting.* Unpublished instrument. (Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203)

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4–15. doi:10.1177/00222194040370010201

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment, 27,* 265–279. doi:10.1177/0734282908330592

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*, 1343–1359. doi:10.1111/j.1467-8624.2007.01069.x

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., . . . Scott, L. (2009). *Assisting students struggling with mathematics: Response to intervention (RTI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Gersten, R., Keating, T., & Irvin, L. K. (1995). The burden of proof: Validity as improvement of instructional practice. *Exceptional Children, 61,* 510–519.

Goldman, S. R., Pellegrino, J. W., & Mertz, D. L. (1988). Extended practice of basic addition facts: Strategy changes in learning disabled students. *Cognition and Instruction, 5*, 223–265. doi:10.1207/s1532690 xci0503_2

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–149). Mahwah, NJ: Erlbaum.

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. 45–58). 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*, 181–195. Retrieved from http:// www.nasponline.org/publications/spr/sprissues.aspx#39

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85,* 103–119. doi:10.1016/S0022-0965 (03)00032-8

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*, 850–867. doi:10.1037 /a0014939

Jordan, N. C., Levine, S. C., & Huttenlocher, J. (1994). Development of calculation abilities in middle- and low-income children after formal instruction in school. *Journal of Applied Developmental Psychology, 15*, 223–240. doi:10.1016/0193-3973(94)90014-0

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

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities, 41*, 451–459. doi: 10.1177/0022219408321126

Mazzocco, M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice, 20*, 142–155. doi: 10.1111/j.1540-5826.2005.00129.x

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review, 37*(3), 359–373. Retrieved from http://www.nasponline.org /publications/spr/index.aspx?vol=37&issue=3

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:10.1177/002221940833 1037

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education. Retrieved from http://www .ed.gov/MathPanel

National Research Council. (2009). *Mathematics learning in early childhood: Paths towards excellence and equity.* Committee on Early Childhood Mathematics, C. T. Cross, T. A. Woods, & H. Schweingruber (Eds.), Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Okamoto, Y., & Case, 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*, 27–59. doi:10.1111 /j.1540-5834.1996.tb00536.x

Pearson Education. (1996). *Stanford achievement test-ninth edition.* Boston, MA: Author.

Pearson Education. (2003). *Stanford achievement test-tenth edition.* Boston, MA: Author.

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

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York, NY: Academic Press.

Scarborough, H. (2001). Connecting early language and literacy to later reading (dis)abilities. In S. B. Neuman, & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 97–110). New York, NY: Guilford.

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children, 77*(1), 37–60.

Siegel, L., Fuchs, L., O'Connor, R., & Vaughn, S. (2011, April). *The future of response to intervention (RTI).* Presentation at the Council for Exceptional Children convention, National Harbor, MD.

Siegler, R. S., & Robinson, M. (1982). The development of numerical understandings. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 241–311). New York, NY: Academic Press.

Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*, 304–325. doi:10.1177/073428290502300402

Swanson, H. L., & Beebe-Frankenberger, M. E. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology, 96*, 471–491. doi:10.1037/0022-0663.96.3.471

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children, 77*, 335–350. doi:10.1260/0014-4029.77.3.335

Wechsler, D. (2003). *Manual for the Wechsler Intelligence Scale for Children-Fourth edition.* San Antonio, TX: The Psychological Corporation.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Tests of Achievement—Revised.* Allen, TX: DLM.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). W*oodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.

ABOUT THE AUTHORS

RUSSELL GERSTEN (California CEC), Executive Director, Instructional Research Group, Los Alamitos, California. **BEN CLARKE** (Oregon CEC), Research Associate, Center on Teaching and Learning, University of Oregon, Eugene. **NANCY C. JORDAN**, Professor, School of Education, University of Delaware, Newark. **REBECCA NEWMAN-GONCHAR,** Senior Research Associate; and **KELLY HAYMOND,** Research Associate, Instructional Research Group, Los Alamitos, California. **CHUCK WILKINS,** Director, Statistics and Evaluation Research, Edvance Research Inc., San Antonio, Texas.

This research was supported in part by Grant Number S283B050034A from the Office of Elementary and Secondary Education of the U.S. Department of Education. The authors wish to thank Sara Woodruff of the University of Georgia for her assistance in the research review and Nicholas Gage of the University of Missouri for his helpful suggestions in refining the manuscript.

Address correspondence concerning this article to Russell Gersten, Executive Director, Instructional Research Group, Los Alamitos, CA (e-mail: rgersten@inresg.org).

Manuscript received December 2010; accepted August 2011.