Exceptional Children

Vol. 77, No. 1, pp. 37-59.
©2010 Council for Exceptional Children.

The Predictive Utility of Kindergarten Screening for Math Difficulty

PAMELA M. SEETHALER

LYNN S. FUCHS
Vanderbilt University

ABSTRACT: *This study examined the reliability, validity, and predictive utility of kindergarten screening for risk for math difficulty (MD). Three screening measures, administered in September and May of kindergarten to 196 students, assessed number sense and computational fluency. Conceptual and procedural outcomes were measured at end of first grade, with MD operationalized as below the 16th percentile. The authors compared single- versus multiple-skill screeners, fall versus spring kindergarten screening, and conceptual versus procedural outcomes. Reliability and validity coefficients were adequate. Logistic regression and receiver operating characteristics analyses indicated that the single- and multiple-skill screeners produced good and similar classification accuracy at the fall and spring screening occasions in forecasting conceptual outcome. To forecast procedural outcome, the screeners produced similar but less accurate fits.*

Prior to the 2004 reauthorization of the Individuals With Disabilities Education Act (IDEA), an IQ-achievement discrepancy was the major approach for identifying learning disability. This identification procedure is problematic for children in kindergarten and first grade, however, because students in the early grades have not had sufficient exposure to academic instruction to accrue a discrepancy. Other possible problems include the "wait-to-fail" nature of this approach (Vaughn & Fuchs, 2003) and the possibility of identifying students as having a learning disability without eliminating poor instructional quality as the reason for poor learning (Vaughn & Fuchs, 2003). A response-to-intervention (RTI) approach represents a major alternative approach to identifying learning disability, as reflected in the 2004 reauthorization.

Implementing evidence-based academic interventions and documenting response to intervention are major features of RTI (Marston, 2005). Students progress through increasingly intensive levels of a prevention system, and only those students for whom standard forms of instruction are deemed insufficient receive formal evaluation for special education services (Fuchs et al., 2007). Although IDEA allows for identification of learning disability within an RTI framework, many questions remain unanswered concerning the implementation of this approach (Mastropieri, & Scruggs, 2005).

Particularly with respect to students in the early grades, screening potential risk for developing learning disability represents an important focus of assessment within RTI. The earlier risk for disability is identified, the sooner efforts can begin to prevent or minimize the effects of that disability. In the area of reading, for example, researchers have documented that poor phonemic awareness and letter-sound knowledge for young students predicts future reading difficulty (e.g., Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004), and early intervention efforts for kindergarten and first-grade students at risk for reading disability have proven effective (e.g., Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Torgesen et al., 1999).

By contrast, a construct or set of skills that represents a strong predictor of future mathematics disability has yet to be identified. A 2005 issue of the *Journal of Learning Disabilities* focused in large part on this problem. Gersten, Jordan, and Flojo (2005) summarized research on early screening for mathematics disability, concluding that a screening instrument for 5- and 6-year-olds based on the skills of counting/simple computation or a sense of quantity/use of mental number lines show promise. These skills are considered aspects of *number sense* (e.g., Dehaene, 1997; Okamoto & Case, 1996), which may serve as a predictor of mathematics performance. As Berch (2005) and Dowker (2005) noted, however, number sense is not clearly defined or easily operationalized. In spite of the ambiguous nature of number sense, screening measures that incorporate aspects of number sense such as counting skill or quantity discrimination may prove useful for forecasting which young students are at risk for mathematics disability (Gersten et al., 2005). In the research literature and this article, *mathematics disability* is operationalized as low mathematics performance and referred to as *mathematics difficulty* (MD).

PRIOR WORK DETERMINING MD RISK OF KINDERGARTEN STUDENTS

We identified 13 studies that targeted kindergarten students, included screening measures or outcome variables specific to mathematics performance, and reported predictive validity or predictive utility: Baker et al., 2002; Bramlett, Rowell, and Mandenberg, 2000; Chard et al., 2005; Clarke, Baker, Smolkowski, and Chard, 2008; Jordan, Kaplan, Locuniak, and Ramineni, 2007; Kurdek and Sinclair, 2001; Lembke and Foegen, 2005; Locuniak and Jordan, 2008; Mazzocco and Thompson, 2005; Pedrotty Bryant, Bryant, Kim, and Gersten, 2006; Simner, 1982; Tiesl, Mazzocco, and Myers, 2001; and VanDerHeyden, Witt, Naquin, and Noell, 2001. Chard et al., Lembke and Foegen, and Pedrotty Bryant et al. included samples of kindergarten and first-grade students; we report results for the kindergarten samples only.

Table 1 lists the number of participants, grades at which screening and outcome assessment took place, screening and outcome measures, correlations between screeners and outcomes, and the predictive utility of measures (i.e., sensitivity, specificity, and overall accuracy, if provided by the authors) for each study. The majority of these studies screened students in kindergarten and assessed mathematics outcome later that same year (Chard et al., 2005; Clarke et al., 2008; Lembke & Foegen, 2005; Pedrotty Bryant et al., 2006; Simner, 1982; VanDerHeyden et al., 2001) or the following year (Baker et al., 2002; Bramlett et al., 2000; Jordan et al., 2007; Simner, 1982; Tiesl et al., 2001). Only four studies (Jordan et al., 2007; Kurdek & Sinclair, 2001; Locuniak & Jordan, 2008; Mazzocco & Thompson, 2005) allowed for greater than 12 months before assessing outcome.

With the exception of Mazzocco and Thompson (2005) and VanDerHeyden et al. (2001), all studies provided data on predictive validity. Correlations ranged from .27 to .72, averaging .51. Five studies provided information on overall accuracy, sensitivity, and specificity, either with predictive validity correlations (Bramlett et al., 2000; Locuniak & Jordan, 2008; Simner, 1982; Tiesl et al., 2001) or without (Mazzocco & Thompson, 2005; VanDerHeyden et al., 2001). For these studies, overall accuracy ranged from 59.8% to 89.4%. Sensitivity ranged widely (00.0%–91.7%), as did specificity (57.5%–94.4%), with the wider spread for sensitivity indicating that screeners were more accurate in

TABLE 1

Predictive Utility of Early Mathematics Screening Studies

						Predictive Validity (r)		*Predictive Utility*		
Study	n	*Grade Screen*^{a}	*Grade Outcome*^{a}	*Screening Measure(s)*	*Outcome Measure(s)*	*(A)*	*(B)*	*Sensitivity*	*Specificity*	*Overall Accuracy*
Baker et al. (2002)	65, 95	K(S)	1(S)	NKT	SAT-9 (A)	.72	.72			
				Digit span backward	NKT (B)	.47	.60			
				Numbers from dictation		.47	.48			
				Magnitude comparison		.54	.45			
Bramlett, Rowell, & Mandenberg (2000)	92	K(F)	1(S)	Informal number probes	WJ-R	.41		75.0%	57.5%	59.8%
Chard et al. (2005)	168	K(F)	K(S)	Count to 20	NKT	.38				
				Count from 6		.39				
				Count from 3		.40				
				Count by 10s		.55				
				Count by 5s		.53				
				Count by 2s		.49				
				Number writing		.57				
				Number identification		.58				
				QD		.50				
				Missing number		.64				
Clarke, Baker, Smolkowski, & Chard (2008)	221	K(F)	K(S)	Oral counting	SESAT	.55				
				Number identification		.58				
				QD		.57				
				Missing number		.60				
Jordan, Kaplan, Locuniak, & Ramineni (2007)	277	K(F)	1(S)	Number sense core	WJ-R Calculations + Applied Problems	.70				
				Counting skills		.36				
				Number knowledge		.54				
				Nonverbal calculation		.52				
				Story problems		.47				
				Number combinations		.58				
Kurdek & Sinclair (2001)	281	K(F)	4(?)	KDI Form Perception	Ohio Grade 4 Achievement Test	.27				
				KDI Number Skills		.37				

continues

TABLE 1 *Continued*

Study	n	*Grade Screen*[a]	*Grade Outcome*[a]	*Screening Measure(s)*	*Outcome Measure(s)*	*Predictive Validity (r)*		*Predictive Utility*		
						(A)	*(B)*	*Sensitivity*	*Specificity*	*Overall Accuracy*
Lembke & Foegen (2005)	44	K(F)	K(S)	QD	Teacher ratings (A)	.64	.33			
				Quantity array	TEMA-3 (B)	.58	.30			
				Number identification		.63	.39			
				Missing number		.44	.41			
Locuniak & Jordan (2008)	198	K(S)	2(W)	Counting	Calculation fluency	.30		51.0%	84.6%	76.3%
				Number knowledge		.45				
				Nonverbal calculations		.51				
				Story problems		.51				
				Number combinations		.57				
Mazzocco & Thompson (2005)	209	K(?)	3(?)	Composite scores from various measures	<10th percentile on TEMA-2 and WJ-R (Calculations)			71.4–91.7%	78.2–90.3%	78.7–89.4%
Pedrotty Bryant, Bryant, Kim, & Gersten (2006)	135	K(W)	K(S)	Oral counting	SAT-10 Math Problem Solving	.49				
				Number identification		.51				
				QD		.61				
				Missing number		.67				
				Digits backward		.54				

Simner (1982)	67
	53
		K(F),
		K(W)
			K(S)
			1(W)
				Writing reversible numbers and letters from short term memory (form errors)	Grade 1 readiness (A)
					Grade 1 report cards (B)
						–.67	–.40	75.0%	89.0%	84.2%

Tiesl, Mazzocco, & Myers (2001)	234	K(S)	1(F–S)	Teacher ratings of math level (<10th percentile)	TEMA-2	.34		65.2%	87.7%	85.4%
VanDerHeyden, Witt, Naquin, & Noell (2001)	25	K(W)	K(S)	Circle number	Retention			71.4	94.4%	88.0%
				Write number	Referral to school team			00.0%	90.9%	80.0%
				Draw circles	“Validation problem”			00.0%	91.7%	88.0%

Note. NKT = Number Knowledge Test (Okamato & Case, 1996); SAT-9 = Stanford Achievement Test, 9th ed. (The Psychological Corporation, 1995); WJ-R = Woodcock-Johnson Psycho-Educational Battery-Revised (Woodcock & Johnson, 1989); QD = Quantity Discrimination (Chard et al., 2005); SESAT = Stanford Early School Achievement Test (Harcourt Educational Measurement, 2003); SAT-10 = Stanford Achievement Test, 10th ed. (Harcourt Assessment Inc., 2003); KDI = Kindergarten Diagnostic Instrument (Robinson & Miller, 1986); TEMA = Test of Early Mathematics Ability (Ginsburg & Baroody, 2003).

[a]S = spring; F = fall; W = winter.

predicting students who would not develop MD than for specifying which students would develop MD.

The majority of studies used single-skill rather than multiple-skill screeners. For example, Bramlett et al. (2000) presented students with randomly ordered numerals between 1 and 20 on a sheet of paper, and students named as many numbers as possible in 1 min. In contrast to the single-skill screening measures, four studies incorporated multiple-skill screeners. Baker et al. (2002), for example, used the Number Knowledge Test (NKT; Okamato & Case, 1996), which samples a range of basic arithmetic concepts and applications. Mazzocco and Thompson (2005) used composite scores from a variety of commercially published tests and subtests of math, reading, and visual-spatial ability to predict future mathematics performance. Across studies, predictive validity was similar for the single- versus multiple-skill screeners. Coefficients for the single-skill screeners ranged from .27 to .67, averaging .50; coefficients for the multiple-skill screeners ranged from .34 to .72, with an average of .51. Although some studies used both types of screeners, none specifically tested which type classified MD with greater precision. This is important because earlier work (e.g., Fuchs et al., 2008) suggested that different aspects of mathematical knowledge may involve separate abilities; a single math task may not effectively predict future math difficulties, as the scope of skills necessary for success expands.

The majority of studies used outcome variables reflecting mathematics performance on published tests (e.g., the Stanford Achievement Test, 9th ed.; The Psychological Corporation, 1995; the Woodcock-Johnson Psycho-Educational Battery-Revised Calculations and Applied Problems subtests, WJ-R, Woodcock & Johnson, 1989). Yet authors also reported outcomes such as teacher ratings (Simner, 1982) and other professional judgments of academic difficulties (VanDerHeyden et al., 2001). Although some of these outcomes related to conceptual understanding of mathematics concepts (e.g., NKT; Okamato & Case, 1996) or procedural outcomes (e.g., WJ-R Calculations subtest; Woodcock & Johnson, 1989), none of the studies addressed whether development could be forecast more precisely for which type of outcome. If aspects of mathematical knowledge involve separate abilities, as some research suggests (Fuchs et al., 2008), then pinpointing the area of difficulty for students yields valuable information for instructional purposes. In terms of decision utility data, the sensitivity of screening variables ranged widely, from 0.00% (i.e., VanDerHeyden et al.'s 2001 prediction of "validation problem") to 91.7% (i.e., Mazzocco & Thompson's 2005 prediction of composite scores on published tests).

The majority of studies assessed mathematical outcomes 1 year or less after initial screening. As a result, students were classified as MD or non-MD before the end of first grade or as early as the kindergarten year. Because kindergarten students vary in their experience with number concepts prior to formal schooling, assessing math outcome before a substantial amount of mathematics instruction occurs potentially yields an inflated estimate of false positives. This is problematic because false positives stress the resources available in schools to provide remediation for students who truly need intervention. Waiting until the students have completed first grade to assess math outcome allows students who have had less preschool exposure to number concepts to "catch up" to peers via formal classroom instruction, and thus should reduce false identification of students as at risk for MD.

Because kindergarten students vary in their experience with number concepts prior to formal schooling, assessing math outcome before a substantial amount of mathematics instruction occurs potentially yields an inflated estimate of false positives.

In addition, the majority of studies relied on predictive correlational data as the sole indication of a measure's ability to forecast MD. Few studies evaluated sensitivity and specificity. This is unfortunate because within RTI, screening tools must accurately pinpoint students with true risk for academic failure so that timely intervention can occur. Therefore, screening studies are tasked to provide evidence of a test's accuracy in classifying students' MD risk. Although predictive correlations can provide general support for the value of kindergarten screening, the decision utility data that attest a screener's value are missing from the majority of previous work.

THE PRESENT STUDY

In the present study, we addressed these limitations in the literature in three ways. First, we adopted a longer perspective than in many prior studies, screening students in the fall and spring of kindergarten and measuring outcomes during the spring of first grade—in effect, a period of two full academic years. Second, in addition to providing evidence of technical adequacy (i.e., reliability; concurrent and predictive validity), we also examined the math screeners' predictive utility (i.e., sensitivity and specificity). Specifically, by holding sensitivity constant at approximately 90%, we were able to consider how the prediction models affected false positives. This allowed us to determine if a single screener or some combination of the screeners resulted in the fewest number of students misidentified as at risk for MD, even while the number of students correctly identified remained constant and high. Third, we extended previous research on the predictive utility of kindergarten math screeners by contrasting (a) the predictive accuracy of single- versus multiple-skill screeners, (b) fall versus spring administration of kindergarten testing, and (c) conceptual versus procedural outcomes.

Our research questions were: What is the reliability of mathematics screening measures for kindergarten students? What are the concurrent and predictive validities with respect to kindergarten and first-grade performance on various mathematics measures? While holding the number of true positives constant (i.e., ~ 90%), which screener or combination of screeners result in the fewest number of false positives? How do single-skill versus multiple-skill math screeners compare in terms of predictive utility? How accurate is fall versus spring kindergarten screening? And, finally, can first-grade mathematics development be forecast more precisely in terms of conceptual or procedural outcomes?

METHOD

PARTICIPANTS

We randomly selected 20 kindergarten teachers from five schools in a southeastern metropolitan school district from a pool of 25 teachers interested in participating. One-half of the kindergarten classrooms received Title I funding. From the 20 classrooms, 252 (of approximately 300) students (i.e., approximately 84%) returned signed parental consent and participated in the initial testing wave in the fall of kindergarten. Of these 252 kindergarten students, 196 completed testing through the end of first grade (or the second year of kindergarten, if retained), an attrition rate of approximately 22% over the 2-year study (i.e., 20 students moved out of the school district before the end of kindergarten, and 36 students moved during first grade). We used inferential statistics to compare students who exited versus those who remained on demographic variables and screening scores. There were no significant differences except on the number sense multiple-skill screener (X = 12.91, *SD* = 6.04 for students who exited the study; 15.65, *SD* = 6.80, for those who remained). We report results for the sample of 196 students with complete data.

Teachers provided demographic information for students and number of minutes of daily math instruction they delivered. The average age of students at the beginning of the study was 5 years 8 months. Of the students, 53% were male, and 52% received free or reduced-price lunch. With respect to race, 36% of the sample was African American, 44% Caucasian, 11% Hispanic, 6% Asian and 3% Kurdish, Indian, Somalian, or Iraqi. Thirteen percent received special education, identified as having a learning disability (< 1%), speech/language disability (6%), giftedness (6%), visual impairment (< 1%), or developmental delay (< 1%). Approximately 5% of students qualified as English language learners; 51% attended preschool prior to kindergarten. As reported by the end of kindergarten, students received an average of 49.08 min of daily math instruction (*SD* = 20.83). The math curriculum used in kindergarten and first grade was *Houghton Mifflin Math* (2004).

Kindergarten Screening Measures

Multiple-Skill Screeners. We created two kindergarten multiple-skill math screeners: Computation Fluency (CF), which is group administered, and Number Sense (NS), which is individually administered. Items were derived from three sources: (a) interviews with experienced kindergarten and first-grade teachers; (b) examination of the existing literature base and the district's kindergarten academic standards; and (c) discussions with university professors familiar with elementary school kindergarten. After piloting the measures with 90 kindergarten students to identify items with poor discrimination, we used WINSTEPS Rasch measurement software (Version 3.58.1) to eliminate or revise items that were inappropriate or ambiguous. We also used the results from the WINSTEPS Rasch software on the individually administered NS measure to order the items by difficulty and derive a ceiling rule for administration, which allowed discontinued testing after five consecutively incorrect answers.

CF is a 5-min timed assessment of counting, addition, and subtraction fluency. It is administered in a whole-class setting and includes 25 items (5 items each of five problem types) presented in random order on one side of an 8 1/2- x 11-inch piece of paper. The five types of items are counting stars in a set, counting two sets of stars, subtracting crossed-out stars from a set, adding arithmetic combinations (presented without star icons), and subtracting arithmetic combinations (without star icons; contact the first author for examples of each problem type). This measure contains five rows of five problems each; the items are bordered in black to help delineate each problem. The student is not penalized for number reversals or poorly formed written responses. Performance is scored as correct responses. We created two forms, identical in format but comprising different items. CF is conceptually based on the Computation CBM probes for Grades 1 through 6 (e.g., Fuchs & Fuchs, 2004; Fuchs, Fuchs, Hamlett, Phillips, & Bentz, 1994). It resembles the Computation CBM probes in appearance, and it samples computation items across the kindergarten curriculum, as do the CBM probes. For analyses, we used the average score across forms at the fall and spring of kindergarten testing occasions.

NS is individually administered. It samples a greater number of mathematics kindergarten skills, with 30 items (3 items each of 10 types) ordered in difficulty from easiest to hardest. The conceptual model for NS is based on early numeracy skills that form the basis of numerical knowledge, that is, knowledge related to counting, patterns, magnitude comparison, and simple arithmetic calculation (Berch, 2005; Gersten et al., 2005; Jordan & Hanich, 2003). Research suggests that deficits in these early numeracy skills may lead to deficient calculation skill and risk for developing MD (Mazzocco & Thompson, 2005). NS comprises items linked to these early numeracy skill areas: three items each of quantity discrimination, mental number lines, ordering numbers, estimation, patterns, counting backwards, shape discrimination, number sentences, writing number sentences, and one-to-one correspondence. For example, for ordering numbers, students are presented with a set of three numbers printed on the page and instructed to write them "in the correct counting order." The first set comprises the numbers 2, 3, 4; the second set, 14, 15, 16; the final set, 18, 19, 20. For quantity discrimination, students are presented with a pair of numbers printed on a page and instructed to circle the larger/smaller item. The student writes answers to items; as with CF, the student is not penalized for misspelled or poorly formed written responses. The score is the number of correctly answered items. Similar to the Concepts/Applications CBM probes (Fuchs, & Fuchs, 2004), NS is a multiple-skill screener that samples within-grade-level skills. However, it differs from the Concepts/Applications CBM probes in that it is not group administered; items are scored immediately after each response; and a ceiling rule limits the length of the test.

Single-Skill Screener. Quantity Discrimination (QD; Chard et al., 2005) is a 1-min timed probe, individually administered, measuring students' ability to name the larger of two numbers (ranging from 0–10), presented in 28 individual boxes across two pages. Clarke et al. (2008) reported test-retest reliability as .85–.99 and concurrent and predictive validity coefficients that ranged from .70 to .80. We chose the QD measure because it has demonstrated strong predictive correlations at first grade (Clarke & Shinn, 2004) and kindergarten (Chard et al., 2005).

Outcome Measures and MD Designation

Math Reasoning and Numerical Operations. The Early Math Diagnostic Assessment (EMDA; The Psychological Corporation, 2002a) is an individually administered norm-referenced test for use with preschool to third-grade students. The test comprises two sections. Math Reasoning measures skills such as counting, ordering numbers, identifying/comparing shapes, problem solving with whole numbers, patterns, time, money, graphs, and measurement. Numerical Operations measures one-to-one correspondence, number identification, number writing, calculation, and rational numbers. Reliability ranges from .71 to .93. Correlations with the Wechsler-Individual Achievement Test (The Psychological Corporation, 2002b) are .82 and .78; with the Wide Range Achievement Test-Revised (Wilkinson, 1993), .67 and .77.

Numeration and Estimation. KeyMath-Revised (KM-R; Connolly, 1998) is an individually administered norm-referenced test for use with students from kindergarten through Grade 12. We used two subtests: Numeration (i.e., concepts such as counting, correspondence, sequencing numbers, and ordinal positions) and Estimation (i.e., estimation of rational numbers, measurement, and computation). Reliability ranges from .50 to .70 for the subtests. Correlations with the Total Mathematics Score of the Iowa Test of Basic Skills (Hoover, Hieronymous, Dunbar, & Frisbie, 1993) and the KM-R Numeration and Estimation subtests are .67 and .43, respectively.

CBM Computation and Concepts/Applications. First-Grade Computation and Concepts/Applications CBM (Fuchs & Fuchs, 2004) sample items from the first-grade curriculum. Students have 3 min to respond to 25 items for Computation; they have 10 min to respond to 22 items for Concepts and Applications. The CBM assessments were scored as number of items correct for Computation and number of missing blanks filled in correctly for Concepts and Applications (e.g., one item on Concepts and Applications requires the student to write the number of tens and the number of ones for a 2-digit numeral; each blank correctly answered yields one point). Although the CBM Computation and Concepts and Applications criterion tests can be administered as whole-class progress-monitoring measures, we did all assessments of final math outcome (including the two CBM tests) individually.

MD Designation. Students received a designation of MD if they scored below the 16th percentile on the EMDA Math Reasoning subtest or the EMDA Numerical Operations subtest (The Psychological Corporation, 2002a) at the end of first grade (or the end of the second year of kindergarten, if a student repeated kindergarten).

Interscorer Agreement and Data-Entry Accuracy

After the first wave of testing (i.e., fall of kindergarten), a second scorer independently scored 20% of protocols. Interscorer agreement (the number of agreed points divided by the total number of points) ranged from 99.29% to 100.0%. This was repeated after the second testing wave (i.e., spring of kindergarten), when interscorer agreement ranged from 98.96% to 100.0%. Interscorer agreement for the fall of kindergarten administration of tests was as follows: QD (Chard et al., 2005), 100.00 %; CF 1, 99.84%; CF 2, 99.60%; NS, 99.73%; KM-R Numeration, (Connolly, 1998) 99.75% ; KM-R Estimation, 100.00%; EMDA Math Reasoning (The Psychological Corporation, 2002a), 100.00%. Interscorer agreement for the spring of kindergarten administration of tests was as follows: QD, 100.00%; CF 1, 100.00%; CF 2, 100.00%; NS, 99.65%; KM-R Numeration, 98.96%; KM-R Estimation, 100.00%; EMDA Math Reasoning, 99.91%. Following the third testing wave (i.e., spring of first grade), 100% of the testing protocols were rescored and entered by a second scorer into a second database. The two databases were compared for discrepancies, which were resolved by examining the original protocols. In this way, we ended up with a single database with no data-entry errors.

Procedure

We administered tests to students in three waves. During the first wave, in fall of kindergarten on three separate days across 3 weeks, students completed one form of CF, NS, both subtests of the EMDA (The Psychological Corporation, 2002a), both subtests of the KM-R (Connolly, 1998), and the alternate form of CF and QD (Chard et al., 2005). (Note: Students completed CF in a whole-class setting the first day; they completed it individually on the third testing day.) During the second testing wave (i.e., spring of kindergarten), students were again tested across 3 weeks and on 3 separate days. The testing schedule was identical to that of the first wave, except that both administrations of CF were conducted in groups.

The third testing wave occurred during the final weeks of the subsequent school year. Again, assessment occurred over 3 weeks and on 3 separate days. At this wave, students completed alternate forms of CBM Computation and CBM Concepts/Applications tests (Fuchs & Fuchs, 2004), the EMDA (The Psychological Corporation, 2002a) subtests, and the KM-R Numeration subtest (Connolly, 1998). (Because of a floor effect for the KM-R Estimation subtest when administered the previous times, we omitted this test from the final testing wave.) All testing in this third wave was conducted individually. Testers were graduate students with varying degrees of classroom experience, who were trained to acceptable levels of accuracy during practice sessions and monitored by the first author throughout all testing waves.

Data Analysis

Reliability of the Screening Measures. To examine the reliability of the kindergarten screening, we evaluated the internal consistency reliability (i.e., coefficient alpha) of both multiple-skill screeners.

Correlations Among Screening and Outcome Measures. We examined the concurrent validity of the three kindergarten screening measures (i.e., QD, Chard et al., 2005; CF, and NS) by correlating the results from the fall and spring administrations with each mathematics outcome measure administered at the same time. Further, we computed Pearson product moment correlation coefficients for the fall administration of the screening measures and the spring administration of the outcome measures to examine the predictive validity from the beginning to the end of kindergarten. To assess predictive validity from the beginning of kindergarten to the end of first grade and from the spring of kindergarten to the end of first grade, we correlated the kindergarten fall and spring screening scores with the first-grade EMDA subtests (The Psychological Corporation, 2002a), KM-R subtest (Connolly, 1998), and CBM mathematics tests.

Classifying Risk for MD. We used logistic regression to evaluate the utility of the kindergarten screening measures for classifying MD status at the end of first grade, separately for math reasoning (i.e., conceptual) and numerical operations (i.e., procedural) outcomes, while holding sensitivity constant. Within the context of RTI, we were interested in maximizing the number of students who truly required additional and intensive mathematics instruction (i.e., students who were identified as at risk for MD at the time of screening and who completed first grade meeting our criterion for MD, or "true positives") while limiting the number of those who did not (i.e., students who were identified as at risk for MD at the time of screening and who completed first grade above our criterion for MD, or "false positives"). The set of true and false positives comprises students identified for secondary intervention. For this reason, we held sensitivity as at least 87.5% of students and then observed how the various models affected specificity. We used SPSS 16.0 statistical software to generate the logistic regression models, entering the screeners in stages to contrast their predictive capabilities.

ROC Curves to Contrast Various Models. We used measures of sensitivity, specificity, overall hit rate, and area under the receiver operating characteristics (ROC) curve (AUC) to contrast the utility of various logistic regression models. *Sensitivity* refers to the proportion of children correctly predicted by the model to be MD in this study and is computed by dividing the number of true positives by the sum of true positives and false negatives. *Specificity*, by contrast, represents the proportion of children correctly predicted to be non MD. Specificity is computed by dividing the number of true negatives by the sum of true negatives and false positives. The overall hit rate is the proportion of children correctly classified as either MD or non MD, and represents the overall accuracy of a prediction model. Finally, the AUC is a plot of the true positive rate against the false positive rate for the different possible cut-points of a test. To contrast the predictive accuracy of logistic regression models, we used the AUC as a measure of discrimination (Swets, 1992). To illustrate, imagine that we had already placed children into their correct MD or non-MD group. If we then selected one child at random from each group, we would assume that the child scoring higher on the kindergarten screeners would be the child from the non-MD group. The AUC, which represents the proportion of randomly chosen pairs of students for which the screeners correctly classified as MD versus non MD, ranges from .50 to 1.00. The greater the AUC, the less likely classification is due to chance. An AUC below .70 indicates a poor predictive model; .70 to .80, fair; .80 to .90, good; and greater than .90, excellent (e.g., Fuchs et al., 2007). The output from ROC analyses includes confidence intervals for the AUC; a lack of overlap for the confidence intervals across models indicates significant difference in predictive accuracy.

RESULTS

Table 2 provides means and standard deviations of each test at each testing wave. For CF, we report the average score for the two forms of the test at the fall and spring kindergarten testing waves. We also report the average score for the CBM tests from the spring of Grade 1 testing wave.

Reliability and Validity of Kindergarten Screening Measures

We were interested in the reliability of only the two multiple-skill screeners because previous work has evaluated the reliability of the single-skill QD measure (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Lembke & Foegen, 2005; Pedrotty Bryant et al., 2006). We evaluated inter-item consistency for the fall administration of CF as follows. Students completed two forms of the measure (i.e., Forms A and B). Half the students were randomly selected to complete Form A during the first (group) administration and Form B during the second (individual) administration; the remaining students completed Form B first and then Form A. We then computed coefficient alpha for the four sets of data and averaged the results. We repeated this procedure in the spring of kindergarten, although at this wave, CF was administered in a group format at both occasions. Alpha for CF averaged .88 for the fall administration and .92 for the spring administration. For the same students, coefficient alpha for NS was .91 for the fall administration and .88 for the spring. All coefficients were significant.

We examined concurrent and predictive validity with various mathematics outcome measures. With respect to concurrent validity, Table 3 provides the zero-order correlations for the fall kindergarten screening and criterion measures below the diagonal and provides the same information for spring of kindergarten above the diagonal. All correlations at both testing occasions were significant. Similar to the concurrent validity correlations, all predictive validity correlations were significant. Tables 4, 5, and 6 provide the zero-order correlations among fall and spring kindergarten measures, fall kindergarten and spring of first-grade measures, and spring of kindergarten and spring of first-grade measures, respectively.

MD Prevalence as a Function of Mathematics Outcome

We determined MD prevalence for students based on their performance on criterion measures administered at the end of first grade, which allowed 2 academic years to elapse from the initial screening occasion to the final measurement of mathematics outcome. MD designation was operationalized as scoring below the 16th percentile on the EMDA Math Reasoning subtest or the EMDA Numerical Operations subtest (The Psychological Corporation, 2002a). The former focused primarily on conceptual skills and mental manipulation of whole numbers (MD-conceptual). The EMDA Numerical Operations subtest measured ability to identify and write numerical symbols and perform written calculations (MD-procedural). Forty students (20.41% of the sample) were MD-conceptual; 59 students (30.10%) were MD-procedural. Twenty-one students (10.71%) met criteria for both MD designations. Chi-square analyses showed no significant differences when comparing the MD groups with their

TABLE 2

Means and Standard Deviations for Kindergarten and Grade 1 Measures

	Grade K Fall				*Grade K Spring*				*Grade 1 Spring*			
Measure	M[a]	(SD)[a]	M[b]	(SD)[b]	M[a]	(SD)[a]	M[b]	(SD)[b]	M[a]	(SD)[a]	M[b]	(SD)[b]
CF1 (K fall: group)	7.55	(5.12)	—	—	16.27	(6.12)	—	—	—	—	—	—
CF2 (K fall: individual)	11.22	(5.72)	—	—	17.58	(6.14)	—	—	—	—	—	—
CF Average	9.38	(5.03)	—	—	16.92	(5.79)	—	—	—	—	—	—
NS	15.65	(6.80)	—	—	21.84	(5.57)	—	—	—	—	—	—
KM-R Num	4.71	(1.90)	103.54	(12.41)	6.39	(2.14)	109.31	(11.62)	9.20	(3.38)	106.76	(13.03)
KM-R Est	1.08	(1.12)	—	—	1.09	(1.42)	—	—	—	—	—	—
EMDA MR	12.42	(4.64)	99.92	(13.55)	17.46	(4.99)	106.68	(14.85)	22.77	(5.66)	98.27	(14.77)
EMDA NO	6.29	(2.01)	101.63	(11.38)	8.14	(1.81)	103.93	(11.99)	10.81	(2.34)	95.01	(14.72)
QD	16.45	(10.13)	—	—	25.89	(10.09)	—	—	—	—	—	—
CBM Comp, Form 1	—	—	—	—	—	—	—	—	12.22	(4.77)	—	—
CBM Comp, Form 2	—	—	—	—	—	—	—	—	12.94	(5.74)	—	—
CBM Comp, average	—	—	—	—	—	—	—	—	12.58	(4.93)	—	—
CBM C/A, Form 1	—	—	—	—	—	—	—	—	21.23	(4.22)	—	—
CMB C/A, Form 2	—	—	—	—	—	—	—	—	20.31	(4.70)	—	—
CBM C/A, average	—	—	—	—	—	—	—	—	20.77	(4.16)	—	—

Note. *n* = 196. CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest (Connolly, 1998); KM-R Est = KeyMath-Revised, Estimation subtest (Connolly, 1998); EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest (The Psychological Corporation, 2002a); EMDA NO = Early Mathematics Diagnostic Assessment, Numerical Operations subtest (The Psychological Corporation, 2002a); QD = Quantity Discrimination; CBM Comp = Grade 1 Curriculum-Based Measurement Computation probe (Fuchs & Fuchs, 2004); CBM C/A = Grade 1 CBM Concepts and Applications probe (Fuchs & Fuchs, 2004).

[a] Raw score. [b] Standard score.

TABLE 3

Concurrent Validity: Correlations Among Kindergarten Screening and Criterion Measures

Measures	*CF1*	*CF2*	*CF Average*	*NS*	*QD*	*KM-R Num*	*KM-R Est*	*EMDA MR*	*EMDA NO*
CF1	—	.79	.94	.67	.61	.62	.35	.71	.64
CF2	.72	—	.95	.69	.64	.60	.34	.68	.62
CF average	.92	.94	—	.72	.66	.64	.36	.74	.67
NS	.58	.67	.68	—	.68	.68	.38	.74	.55
QD	.55	.67	.66	.71	—	.61	.34	.64	.56
KM-R Num	.55	.59	.62	.67	.64	—	.41	.68	.58
KM-R Est	.26	.29	.30	.30	.31	.32	—	.49	.40
EMDA MR	.60	.68	.69	.79	.66	.67	.39	—	.66
EMDA NO	.56	.59	.62	.68	.60	.61	.26	.62	—

Note. Values below the diagonal correspond to correlations for fall kindergarten screening and criterion measures; values above the diagonal correspond to correlations for spring kindergarten screening and criterion measures. All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest (Connolly, 1998); KM-R Est = KeyMath Revised, Estimation subtest (Connolly, 1998); EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest (The Psychological Corporation, 2002a); EMDA NO = Early Mathematics Diagnostic Assessment, Numerical Operations subtest; (The Psychological Corporation, 2002a) QD = Quantity Discrimination.

non-MD counterparts on gender, socioeconomic status, race, special education status, English language learner status, or number of daily minutes of math instruction, except as follows. On gender, the MD-procedural group comprised twice as many boys (68%) as girls (32%). On socio-economic status, the MD-conceptual group comprised twice as many students receiving free or reduced lunch (68%) as those who did not (32%).

TABLE 4

Predictive Validity: Correlations Among Kindergarten Screening, Fall and Spring

	Spring Kindergarten Screening								
Fall Kindergarten	*CF1*	*CF2*	*CF Average*	*NS*	*QD*	*KM-R Num*	*KM-R Est*	*EMDA MR*	*EMDA NO*
CF1	.58	.52	.58	.54	.58	.48	.61	.51	.49
CF2	.67	.62	.67	.64	.66	.44	.68	.57	.62
CF average	.67	.62	.68	.64	.67	.49	.70	.58	.60
NS	.68	.63	.69	.82	.71	.40	.74	.56	.62
QD	.64	.64	.68	.71	.68	.34	.65	.53	.75

Note. All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest (Connolly, 1998); KM-R Est = KeyMath-Revised, Estimation subtest (Connolly, 1998); EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest (The Psychological Corporation, 2002a); EMDA NO = Early Mathematics Diagnostic Assessment, Numerical Operations subtest (The Psychological Corporation, 2002a); QD = Quantity Discrimination.

TABLE 5

Predictive Validity: Correlations Among Fall Kindergarten Screening and Spring Grade 1 Measures

	Spring Grade 1								
Fall Kindergarten	*KM-R Num*	*EMDA MR*	*EMDA NO*	*CBM1*	*CBM2*	*CBM Average*	*C/A1*	*C/A2*	*C/A Average*
CF1	.58	.59	.56	.41	.45	.46	.42	.44	.46
CF2	.64	.65	.53	.45	.48	.50	.50	.50	.54
CF average	.66	.67	.58	.46	.50	.52	.50	.51	.54
NS	.72	.70	.55	.48	.55	.56	.62	.63	.67
QD	.65	.66	.52	.43	.56	.53	.52	.56	.58

Note. All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest (Connolly, 1998); EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest (The Psychological Corporation, 2002a); EMDA NO = Early Mathematics Diagnostic Assessment, Numerical Operations subtest (The Psychological Corporation, 2002a); CBM1 = Grade 1 Curriculum-Based Measurement Computation probe (Fuchs & Fuchs, 2004), first administration; CBM2 = second administration; C/A1 = Grade 1 Concepts and Applications probe (Fuchs & Fuchs, 2004), first administration; C/A2 = second administration; QD = Quantity Discrimination.

ROC Curves to Contrast the Predictive Utility of Logistic Regression Models

In Table 7, we report results of the logistic regression analyses for classifying MD status at the end of first grade, with respect to conceptual and procedural outcomes. The table shows the hit rate (i.e., overall accuracy), sensitivity, specificity, and area under the ROC curve (AUC) of the three kindergarten math screeners when administered to students in the fall and in the spring.

For classifying MD-conceptual based on the fall-administered screeners (see the top half of Table 7), while holding sensitivity high (87.5%–90.0%), the single-skill QD measure (Chard et

TABLE 6

Predictive Validity: Correlations Among Spring Kindergarten Screening and Spring Grade 1 Measures

	Spring Grade 1								
Spring Kindergarten	*KM-R Num*	*EMDA MR*	*EMDA NO*	*CBM1*	*CBM2*	*CBM Average*	*C/A1*	*C/A2*	*C/A Average*
CF1	.60	.66	.59	.51	.53	.56	.55	.58	.61
CF2	.59	.62	.51	.45	.51	.52	.53	.55	.58
CF average	.63	.68	.58	.51	.55	.57	.57	.60	.63
NS	.70	.72	.55	.48	.56	.56	.66	.68	.72
QD	.62	.62	.47	.44	.54	.53	.49	.54	.55

Note. All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest (Connolly, 1998); EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest (The Psychological Corporation, 2002a); EMDA NO = Early Mathematics Diagnostic Assessment, Numerical Operations subtest (The Psychological Corporation, 2002a); CBM1 = Grade 1 Curriculum-Based measurement Computation probe (Fuchs & Fuchs, 2004), first administration; CBM2 = second administration; C/A1 = Grade 1 Concepts and Applications probe (Fuchs & Fuchs, 2004), first administration; C/A2 = second administration; QD = Quantity Discrimination.

TABLE 7

Classification Indices for Logistic Regression Models

									Hit			*ROC*		
Outcome/Model	B	SE	*Wald*	p	*TN*	*FN*	*TP*	*FP*	*Rate*	*Sensitivity*	*Specificity*	*AUC*	SE	*CI*
MD-Conceptual: Fall predictors														
Quantity discrimination	–.206	.037	30.233	.000	103	4	36	53	70.9	90.0	66.0	.857	.030	.797–.916
Constant	1.042	.386	7.288	.007										
Computation fluency	–.333	.066	25.061	.000	89	5	35	67	63.3	87.5	57.1	.802	.034	.735–.868
Constant	1.140	.463	6.054	.014										
Number sense	–.207	.035	35.007	.000	100	4	36	56	69.4	90.0	64.1	.841	.030	.783–.900
Constant	1.377	.446	9.525	.002										
MD-Conceptual: Spring predictors														
Quantity discrimination	–.168	.026	40.187	.000	104	4	36	52	71.4	90.0	66.7	.861	.035	.793–.929
Constant	2.303	.548	17.649	.000										
Computation fluency	–.276	.045	37.657	.000	109	4	36	47	74.0	90.0	69.9	.860	.028	.806–.915
Constant	2.655	.611	18.890	.000										
Number sense	–.355	.051	37.416	.000	118	5	35	38	78.1	87.5	75.6	.877	.028	.822–.931
Constant	4.887	.986	24.544	.000										
MD-Procedural: Fall predictors														
Quantity discrimination	–.074	.018	16.314	.000	44	6	53	93	49.5	89.8	32.1	.690	.040	.612–.768
Constant	.268	.298	.808	.369										
Computation fluency	–.151	.040	14.081	.000	49	5	54	88	52.6	91.5	35.8	.672	.041	.592–.752
Constant	.459	.359	1.641	.200										
Number sense	–.110	.025	19.028	.000	44	7	52	93	49.0	88.1	32.1	.696	.040	.619–.774
Constant	.775	.388	3.987	.046										
MD-Procedural: Spring predictors														
Quantity discrimination	–.062	.017	14.105	.000	33	7	52	104	43.4	88.1	24.1	.661	.043	.577–.745
Constant	.701	.426	2.705	.100										
Computation fluency	–.136	.030	21.114	.000	56	6	53	81	55.6	89.8	40.9	.722	.037	.649–.794
Constant	1.343	.484	7.703	.006										
Number sense	–.130	.030	18.164	.000	39	7	52	98	46.4	88.1	28.5	.687	.041	.605–.768
Constant	1.914	.655	8.551	.003										

Note. MD = math difficulty; TN = true negatives; FN = false negatives; TP = true positives; FP = false positives; ROC = receiver operating characteristics; AUC = area under ROC curve.

al., 2005) resulted in specificity of 66.0%; the hit rate was 70.9%. The multiple-skill screeners, CF and NS, produced similar results. Specificity for those screeners were 57.1% and 64.1%, respectively; hit rates were 71.9% and 78.1%. The AUCs for the three fall screeners were .857, .802, and .841, which are deemed good (Fuchs et al., 2007). Confidence intervals for the AUCs overlapped, indicating that the models were not significantly different. Based on the fall screeners, 4 to 5 students who were designated MD-conceptual were missed (i.e., see FN column); 53 to 67 students identified as at risk did not the meet end-of-first-grade criterion for MD-conceptual (i.e., see FP column).

For predicting the same MD-conceptual outcome but based on the spring-administered screening measures (see middle of Table 7), again holding sensitivity at ~ 90.0%, results were similar. The single-skill and multiple-skill screeners resulted in specificity of 66.7% (QD; Chard et al., 2005), 69.9% (CF), and 75.6% (NS); hit rates were 71.4%, 74.0%, and 78.1% for the screeners, respectively. AUCs ranged from .860 to .877, which are deemed good, and overlapping confidence intervals again attested to the comparability of the models. Although the number of false negatives remained the same (i.e., 4 to 5), the number of false positives was lower with the spring administration of CF (from 67 to 47), and of NS (from 56 to 38). The number of false negatives for QD decreased by only one in the spring (from 53 to 52).

For predicting MD-procedural status, the three screeners performed similarly in the fall and in the spring (see bottom half of Table 7). Holding sensitivity at ~ 90.0%, specificity for QD (Chard et al., 2005), CF, and NS based on fall screening was 32.1%, 35.8%, and 32.1%, respectively. Based on spring screening, specificity was similar: 24.1%, 40.9%, and 28.5%, respectively. Hit rates across both testing occasions ranged from 43.4% to 55.6%. With the exception of the spring-administered CF, which resulted in an AUC of .722 (deemed fair), the screeners' AUCs were all less than .70 (deemed poor). Holding the number of false negatives low at 5 to 7, false positives ranged from 81 to 104. With the exception of CF, with which the number of false positives decreased from the fall to spring screening (by 7 students), the number of false positives increased in the spring. Based on overlapping confidence intervals of the AUCs, the predictive utility of the three screening measures were comparable at both kindergarten screening occasions.

Although there were no significant differences when looking separately at MD-conceptual and MD-procedural results (i.e., screeners performed similarly, irrespective of testing occasion, when predicting MD-conceptual or MD-procedural status), there was a significant difference when combining results. That is, the screeners classified future MD status in terms of conceptual outcome with significantly greater accuracy than in terms of procedural outcome. The AUCs for the three screeners when predicting MD-conceptual were higher than when predicting MD-procedural; their nonoverlapping confidence intervals indicated statistically significant fits.

In Table 8, we report the results of the logistic regression analyses for various combinations of the screeners, again while holding sensitivity at ~90.0%. Models comprising combinations of the fall administration of the screeners for predicting MD-conceptual yielded specificity ranging from 66.0% to 70.5% and hit rates of 71.9% to 74.5%. Keeping the number of false negatives at 4, the number of false positives ranged from 46 to 51 for the various models. AUCs ranged from .856 to .878, all deemed good, and did not overlap, indicating statistical comparability. Prediction models comprising combinations of the spring administration of the screeners, although not statistically different from the fall models, resulted in fewer false positives (29 vs. 45). Specificity from the spring models ranged from 71.2% to 81.4%; hit rates from 75.0% to 83.2% AUCs for each of the spring prediction models approximated .90, which is deemed excellent.

Models predicting MD-procedural based on combinations of the fall administration of the screeners, keeping sensitivity constant at ~ 90.0%, resulted in specificity ranging from 31.4% to 38.7% and hit rates of 48.5% to 54.1%. AUCs ranged from .702 to .713, all deemed fair; nonoverlapping confidence intervals of the models indicated statistical comparability. Models predicting MD-procedural based on combinations of the spring screeners were comparable. The models' specificity ranged from 28.5% to 43.8% and

TABLE 8

Classification Indices for Logistic Regression Models

									Hit			ROC		
Outcome/Model	B	SE	*Wald*	p	*TN*	*FN*	*TP*	*FP*	*Rate*	*Sensitivity*	*Specificity*	*AUC*	*SE*	*CI*
MD Conceptual: Fall Predictors														
Quantity Discrimination	–.165	.041	16.637	.000	105	4	36	51	71.9	90.0	67.3	.871	.029	.815–.927
Computation Fluency	–.164	.078	4.446	.035										
Constant	1.780	.539	10.912	.001										
Quantity Discrimination	–.150	.042	12.756	.000	106	4	36	50	72.4	90.0	67.9	.871	.028	.816–.926
Number Sense	–.105	.044	5.709	.017										
Constant	1.761	.503	12.259	.000										
Computation Fluency	–.208	.077	7.318	.007	103	4	36	53	70.9	90.0	66.0	.856	.029	.800–.912
Number Sense	–.153	.040	14.696	.000										
Constant	2.166	.573	14.287	.000										
Quantity Discrimination	–.130	.044	8.950	.003	110	4	36	46	74.5	90.0	70.5	.878	.022	.826–.931
Computation Fluency	–.126	.082	2.382	.123										
Number Sense	–.086	.046	3.453	.063										
Constant	2.201	.601	13.430	.000										
MD Conceptual: Spring Predictors														
Quantity Discrimination	–.115	.030	14.491	.000	120	4	36	36	79.6	90.0	76.9	.890	.030	.832–.948
Computation Fluency	–.172	.052	10.945	.001										
Constant	3.614	.723	25.022	.000										
Quantity Discrimination	–.103	.031	11.355	.001	120	4	36	36	79.6	90.0	76.9	.893	.029	.837–.949
Number Sense	–.216	.058	13.763	.000										
Constant	5.119	1.033	24.548	.000										
Computation Fluency	–.170	.054	9.838	.002	111	4	36	45	75.0	90.0	71.2	.898	.025	.849–.947
Number Sense	–.225	.058	15.263	.000										
Constant	5.497	1.069	26.460	.000										
Quantity Discrimination	–.081	.033	5.985	.014	127	4	36	29	83.2	90.0	81.4	.906	.026	.856–.957
Computation Fluency	–.118	.058	4.100	.043										
Number Sense	–.175	.062	7.996	.005										
Constant	5.485	1.084	25.620	.000										

MD Procedural: Fall Predictors														
Quantity Discrimination	–.049	.023	4.591	.032	42	6	53	95	48.5	89.8	30.7	.702	.040	.623–.781
Computation Fluency	–.085	.049	2.969	.085										
Constant	.625	.368	2.888	.089										
Quantity Discrimination	–.038	.025	2.461	.117	53	6	53	84	54.1	89.8	38.7	.710	.039	.634–.787
Number Sense	–.072	.034	4.486	.034										
Constant	.798	.391	4.177	.041										
Computation Fluency	–.077	.050	2.406	.121	43	6	53	94	49.0	89.8	31.4	.708	.040	.629–.786
Number Sense	–.078	.032	6.076	.014										
Constant	.973	.417	5.459	.019										
Quantity Discrimination	–.029	.026	1.204	.273	50	6	53	87	52.6	89.8	36.5	.713	.040	.635–.790
Computation Fluency	–.057	.053	1.168	.280										
Number Sense	–.059	.036	2.619	.106										
Constant	.939	.417	5.064	.024										
MD Procedural: Spring Predictors														
Quantity Discrimination	–.020	.022	.874	.350	56	6	53	81	55.6	89.8	40.9	.722	.037	.649–.795
Computation Fluency	–.112	.038	8.662	.003										
Constant	1.473	.506	8.481	.004										
Quantity Discrimination	–.025	.022	1.335	.248	39	6	53	98	46.9	89.8	28.5	.694	.041	.613–.775
Number Sense	–.100	.040	6.264	.012										
Constant	1.896	.651	8.476	.004										
Computation Fluency	–.093	.040	5.566	.018	60	6	53	77	57.7	89.8	43.8	.728	.037	.655–.801
Number Sense	–.065	.040	2.557	.110										
Constant	2.027	.656	9.549	.002										
Quantity Discrimination	–.007	.024	.091	.763	58	6	53	79	56.6	89.8	42.3	.728	.037	.655–.801
Computation Fluency	–.088	.043	4.322	.038										
Number Sense	–.059	.044	1.800	.180										
Constant	2.015	.656	9.428	.002										

Note. MD = math difficulty; TN = true negatives; FN = false negatives; TP = true positives; FP = false positives; ROC = receiver operating characteristics; AUC = area under ROC curve.

hit rates ranged from 46.9% to 57.7%. The range of AUCs for the spring models (i.e., .694 to .728) were not significantly different than for the fall models, although false positives decreased by 14 for the combination of QD (Chard et al., 2005) and CF and by 17 for the combination of CF and NS.

Based on these analyses, the model yielding the fewest false positives (46) with respect to MD-conceptual, based on fall of kindergarten testing, comprised all three screeners. Predicting the same MD-conceptual status, yet based on spring of kindergarten testing, the three screeners in combination again resulted in the fewest false positives (29). To predict MD-procedural status from fall of kindergarten testing, the combination of QD (Chard et al., 2005) and NS resulted in the fewest false positives (84). In the spring of kindergarten, that number decreased to 77 when using CF and NS for predicting.

DISCUSSION

We evaluated the technical adequacy and predictive utility of one single-skill and two multiple-skill measures for screening kindergarten students for risk for MD. The single-skill screener assessed students' ability to quickly discriminate magnitudes within pairs of numerals ranging from 0 to 10. The multiple-skill screeners assessed computational fluency and various mathematical concepts central to early mathematical development. Conceptual and procedural math outcomes were assessed at the end of first grade, with MD operationalized as performance below the 16th percentile. Holding sensitivity constant and high (~ 90.0%), we examined changes in specificity and number of false positives for the screeners used singly and in combination.

Performance cutoff scores for determining MD status range widely in the research literature, as Murphy, Mazzocco, Hanich, and Early (2007) noted. Reasons for using a more restrictive cutoff (e.g., < 10th percentile) or a more lenient cutoff (e.g., < 35th percentile) may include, for example, the need to obtain a particular sample size or to reflect school eligibility decision guidelines. Although the designation of a performance cutoff criterion to designate MD status is arbitrary, it is used in research programs by necessity due to the lack of a standard definition of MD. We operationalized MD status as performance below the 16th percentile, rather than using a higher cutoff, in an attempt to identify students with serious math deficits.

Previous studies had investigated the reliability and validity of the single-skill QD screener (Chard et al., 2005; Clarke & Shinn, 2004; Lembke & Foegen, 2005; Pedrotty Bryant et al., 2006). Results from these studies showed reliability to average ~ .90, with concurrent and predictive validity averaging approximately .60. Our results echo these findings with respect to validity. Validity correlations ranged from .57 to .63 with criterion measures (i.e., excluding the KM-R Estimation scores [Connolly, 1998], which demonstrated a floor effect for kindergarten students at the fall and spring testing occasions, resulting in lower correlations). In the present study, we focused our attention on the technical adequacy of the two multiple-skill kindergarten screeners (i.e., CF and NS), even as we considered the validity of the single-skill QD test.

Reliability averages of the two multiple-skill screeners were comparable to what had been found previously for the single-skill screener (i.e., .90 for both the fall and spring administrations); these reliability estimates are in the acceptable range (Urbina, 2004). In terms of concurrent and predictive validity, figures for the multiple-skill screeners were similar to those of the single-skill screener, QD (Chard et al., 2005). For example, with respect to fall-of-kindergarten to end-of-first-grade predictive validity, coefficients ranged from .55 to .72 for the two multiple-skill math screeners with outcome measures versus .52 to .66 for QD. Interestingly, the average predictive validity data for our three math screeners with respect to end-of-first grade math skill remained nearly the same from the fall to the spring testing occasions (i.e., .63 and .62, respectively). These validity estimates for the multiple-skill screeners are higher than the average predictive validity of the kindergarten screening literature, which comprises an assortment of screening and outcome measures. Although some studies showed higher predictive validity correlations (e.g., .72 in Baker et al., 2002; .70 in Jordan et al., 2007), kindergarten math screeners from earlier studies correlated (on average) .50 with future measures of mathematical performance. Because kindergarten students begin school in the fall with varying levels of developmental maturity, attention, or experience with paper-and-pencil tasks, it would be understandable if the relations among math screeners and criterion measures were stronger in the spring, once some of the variability due to unequal preschool experiences is eliminated. Our results did not demonstrate this, however. Predictive validity remained stable across the kindergarten school year, with respect to end-of-first-grade mathematics outcomes.

Number Sense comprises items linked to these early numeracy skill areas: three items each of quantity discrimination, mental number lines, ordering numbers, estimation, patterns, counting backwards, shape discrimination, number sentences, writing number sentences, and one-to-one correspondence. By combining the scores from the items of each skill area, we were able to examine the correlations between each skill area and the first-grade math outcome measures. Correlations for the fall administration of NS with EMDA Math Reasoning (The Psychological Corporation, 2002a) ranged from .40 (shape discrimination) to .61 (mental number lines); for the spring administration of the screener, correlations ranged from .26 (shape discrimination) to .63 (mental number lines). Correlations with Numbers Sense and EMDA Numerical Operations ranged from .28 (shape discrimination) to .48 (counting backwards) for the fall and from .23 (patterns) to .48 (writing numbers) for the spring.

In addition to examining technical aspects of the kindergarten math screeners, we specifically questioned whether the predictive utility of our tests differed as a function of item composition (i.e., single- vs. multiple-skill); the time of year screening occurred (i.e., fall vs. spring of kindergarten); or the focus of mathematical outcome (i.e., conceptual vs. procedural). To our knowledge, no previous work has addressed these issues. If educators are to accurately pinpoint students in need of intensive math intervention, research should inform the practice of *how*, *when*, and *with respect to what outcome* this may best be accomplished.

With respect to *how*, we asked: Might a brief single-skill test of magnitude comparison forecast future math ability of kindergarten students just as well as, or perhaps better than, multiple-skill tests of varied early numerical concepts? To answer this question, we compared the AUCs of the single-skill QD (Chard et al., 2005) to the multiple-skill screeners, at both the fall and spring kindergarten screening occasions, and with respect to two mathematical outcomes at the end of first grade. We found no significant differences in predictive utility for QD versus multiple-skill screening, at fall or spring for either math outcome, indicating that a brief, individually administered measure of quantity discrimination is comparable to the multiple-skill screeners (which include more widely varied arithmetical and numerical items and take slightly longer to administer) in forecasting MD. This is likely welcome news for kindergarten teachers who often have limited time and resources for screening. Of course, separate from the issue of efficiency, the multiple-skill screeners may provide teachers with better information for instructional planning than does QD. Sampling a wider variety of early mathematical skills, as the multiple-skill screeners do, provides an opportunity for identifying students' specific numerical strengths and weaknesses. It also may avoid a ceiling effect during progress monitoring, as has been shown to occur at first grade with the use of a single screener (Fuchs et al., 2007), and future research should contrast the use of multiple- versus single-skill screeners for progress monitoring at kindergarten.

Might a brief single-skill test of magnitude comparison forecast future math ability of kindergarten students just as well as, or perhaps better than, multiple-skill tests of varied early numerical concepts?

Second, in terms of *when*, we asked: Do marked differences exist in decision-making utility when screening students in the fall versus the spring of kindergarten? This question is important to answer for two related and competing reasons. On the one hand, studies show that screening for future reading disability at an early age produces a high proportion of false positives (Catts, 1991; Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009), stressing schools to provide intervention to students who do not require that help. Thus, waiting a few months or even until the kindergarten year is complete may better identify students whose initially low performance during screening results from developmental or experiential lag rather than from risk for MD. If this were the case, one would expect a significant difference in predictive accuracy for fall versus spring screening. On the other hand, refraining from screening students until the spring of kindergarten (or even later), with the belief that fall screening is untrustworthy, denies students months of intervention that may serve to offset or prevent math deficits. To address this dilemma, we compared the AUCs of the fall versus the spring math screeners with respect to the same two end-of-first grade mathematical outcomes. Our results showed no statistically significant differences in predictive utility between the fall and spring screening occasions, underscoring the potential value of beginning early, in the fall of the kindergarten year. Nevertheless, the large numbers of false positives (ranging from 46–67 for conceptual outcomes and from 84–95 for procedural outcomes) suggest that delaying screening until the end of or after first grade may be prudent. This issue should be pursued in future work.

Third, *with respect to what outcome*, we asked: What should we look for in terms of MD? Should educators consider conceptual mathematical deficits as a hallmark of MD at the end of first grade or should the focus be on procedural deficits? Prior work shows that elementary-aged students with MD develop marked deficits in computational fluency as well as number processing (e.g., Jordan, Hanich, & Kaplan, 2003; Mazzocco, 2007). Yet it is plausible that students as young as first graders may simply have insufficient formal experience with paper-and-pencil tasks comprising addition or subtraction facts, such as the EMDA Numerical Operations subtest (The Psychological Corporation, 2002a). Therefore, a math outcome for designating MD that focuses on procedural skill for students at this young age (e.g., solving written number combinations or two-digit addition and subtraction items) may prove less useful than one that focuses on the numeracy concepts more likely to have been taught across kindergarten and first grade, as well as informally in the home environment (e.g., shape identification or the meaning of "more" or "less"). Our results supported this. When we contrasted predictive models with conceptual versus procedural mathematical outcomes, we found those with conceptual outcomes to be more accurate than those with procedural outcomes, regardless of type of screener (i.e., single- or multiple-skill) or time of screening (i.e., fall or spring). During the fall or spring of kindergarten, AUCs for our screening models ranged from .80 to .91 for the EMDA Math Reasoning subtest, indicating good predictive utility for conceptual outcome. By contrast, during the same timeframes, AUCs ranged only from .66 to .73 for the EMDA Numerical Operations subtest, indicating poor predictive utility for procedural outcome. This suggests that we can predict first-grade procedural deficits less accurately than conceptual deficits, at least when screening learners in their kindergarten year.

Educational Implications

In summary, single-skill and multiple-skill screening measures produced good and similar fits at both fall and spring of kindergarten, in terms of forecasting conceptual mathematics outcome at the end of first grade. The practical implications of this for kindergarten teachers are that a brief, individually administered measure of quantity discrimination is comparable to multiple-skill screeners (which include more widely varied arithmetical and numerical items and take slightly longer to administer) in forecasting MD, and that fall testing would allow for a greater length of intervention time for students who fail the screen. Yet, with respect to procedural outcome, the single- and multiple-skill screeners produced similar but significantly less accurate fits. Although our results lend tentative support to the potential of screening students at the beginning of kindergarten for end-of-first-grade MD, additional study is needed to increase overall classification accuracy. Regardless of the predictive model, we found an unacceptably high proportion of students misidentified as false positives. This weakens the decision-making utility of the screeners and raises concerns about one-time universal screening within an RTI framework. Similar findings are accruing in reading (e.g., Jenkins, Hudson, & Johnson, 2007), suggesting the need for a multiple-gating screening procedure, in which (a) a cut-point on the universal screen is set to minimize false negatives, and (b) a more thorough conventional assessment, dynamic assessment, or short-term progress monitoring is conducted among the subset of students who fail the universal screen to discriminate true from false positives.

To illustrate this point, consider the results of our prediction models comprising the screeners used singly (see Table 7) and in combination (see Table 8). Holding sensitivity at a level that minimizes the number of students missed by the screening event (i.e., approximately 10% of the truly at-risk students), we note the unacceptably high numbers of students misidentified as at risk for developing MD, even with the models that yield the fewest numbers of false positives. For example, the prediction model that includes all three screeners administered in the spring of kindergarten and with respect to MD-conceptual yields an overall hit rate of 83.2%. However, this model also incorrectly categorizes 29 students as at risk for MD—even with a lengthened screening event, because all three tests would be administered. Although some models result in fewer numbers of false positives than others, students misidentified as at risk for MD and in need of intensive intervention stress school resources unnecessarily in terms of personnel, materials, and instructional time. Future work should investigate the potential of multiple-gating kindergarten screening procedures to identify risk of MD more precisely. In a more general sense, results show that even with acceptable levels of predictive validity correlations, problems with predictive utility may occur, indicating the importance of incorporating classification analysis.

Limitations

As readers interpret findings, they should consider five study limitations. Three pertain to the participants; two to the nature of the screening measures. First, participants were selected from only one school district in a southeastern metropolitan area. Sampling students from a more diverse and representative population would provide for greater generalizability. Second, although our attrition rate was within reason (22%) given the 2-year timeframe of the study, we note that students who remained through the end of first grade scored higher than those who exited on the fall kindergarten multiple-skill NS screening measure. This raises questions about whether results would change if the 56 children who moved prior to the end of first grade had remained. Even so, the students who exited and those who remained were demographically comparable, and they were mathematically comparable as indexed on the other two screeners. Third, consented students represented less than half of the classroom population, questioning whether results would remain stable had more families/students agreed to participate. Fourth, with respect to the screening measures, we did not address the issue of timed testing in this study. The single-skill QD screener (Chard et al., 2005) and the multiple-skill CF screener were timed; the multiple-skill NS screener was not. Additionally, neither subtest used to determine MD status was timed. Students were aware when they were completing assessments with timed limits, and for some students, timing may have been a distraction or a stressor. Yet, as shown with some reading tests (e.g., Fuchs, Fuchs, Hosp, & Jenkins, 2001), fluency may be an important way of distinguishing students' skill levels, abilities, and potential. In any case, we cannot state whether timed tests make a difference in predictive utility for students at this age. Finally, we did not include measures of general intelligence or reading ability, so we do not know the extent to which these factors may influence predictive utility. Future work should examine the discriminant validity of the screening measures with respect to IQ and reading.

REFERENCES

Baker, S., Gersten, R., Flojo, J., Katz, R., Chard, D. J., & 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.

Bramlett, R. K., Rowell, R. K., & Mandenberg, K. (2000). Predicting first grade achievement from kindergarten screening measures: A comparison of child and family predictors. *Research in the Schools, 7*, 1–9.

Catts, H. (1991). Early identification of dyslexia: Evidence from a follow-up study of speech-language impaired children. *Annals of Dyslexia, 41*, 163–177.

Catts, H.W., Petscher, Y., Schatschneider, C, Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*, 163–177.

Chard, D., Clarke, B., Baker, B., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment Issues in Special Education, 30*, 3–14.

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.

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.

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

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York, NY: Oxford University Press.

Dowker, A. (2005). Early identification and intervention for students with mathematics difficulty. *Journal of Learning Disabilities, 38*, 324–332.

Foorman, B. F., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk-children. *Journal of Educational Psychology, 90*, 37–55.

Fuchs, L.S., & Fuchs, D. (2004). *Using CBM for progress monitoring in math.* Retrieved from http://www.studentprogress.org/summer_institute/2007/Written/WrittenExpression_Handouts_2007.doc

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., Hamlett, C. L., Phillips, N. B., & Bentz, J. (1994). Classwide curriculum-based measurement: Helping general educators meet the challenge of student diversity. *Exceptional Children, 60*, 518–537.

Fuchs, L. S., Fuchs, D, Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.

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

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

Ginsburg, H., & Baroody, A. (2003). *Test of early mathematics ability* (2nd ed.). Austin, TX: PRO-ED.

Harcourt Assessment, Inc. (2003). *Stanford achievement test* (10th ed.). San Antonio, TX: Author.

Harcourt Educational Measurement. (2003). *Stanford early school achievement test.* San Antonio, TX: Author.

Hoover, H. D., Hieronymous, A. N., Dunbar, S. B., & Frisbie, D. A. (1993). *Iowa test of basic skills.* Itasca, IL: Riverside.

Houghton Mifflin math. (2004). Boston, MA: Houghton Mifflin Harcourt.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600.

Jordan, N. C., & Hanich, L. B. (2003). Characteristics of children with moderate mathematical deficiencies: A longitudinal perspective. *Learning Disabilities Research & Practice, 18*, 213–221.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development, 74*, 834–850.

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.

Kurdek, L. A., & Sinclair, R. J. (2001). Predicting reading and mathematics achievement in fourth-grade children from kindergarten readiness scores. *Journal of Educational Psychology, 93*, 451–455.

Lembke, E., & Foegen, A. (2005, February). *Monitoring student progress in early math.* Paper presented at the Pacific Coast Research Conference, San Diego, CA.

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

Marston, D. (2005). Tiers of intervention in responsiveness to intervention: Prevention outcomes and learning disabilities identification patterns. *Journal of Learning Disabilities, 38*, 539–544.

Mastropieri, M. A., & Scruggs, T. E. (2005). Feasibility and consequences of response to intervention: Examination of the issues and scientific evidence as a model for the identification of individuals with learning disabilities. *Journal of Learning Disabilities, 38*, 525–531.

Mazzocco, M. M. (2007). Defining and differentiating mathematical learning disabilities. In D. B. Berch & M. M. Mazzocco (Eds.), *Why is math so hard for some children?* (pp. 29–48). Baltimore, MD: Brookes.

Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*, 142–155.

Murphy, M. M., Mazzocco, M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities, 40*, 458–478.

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. In R. Case & Y. Okamoto (Eds.), *The role of central conceptual structures in the development of children's thought: Monographs of the Society for Research in Child Development* (Vol. 1–2, pp. 27–58). Malden, MA: Blackwell.

Pedrotty Bryant, D., Bryant, B. R., Kim, S. A, & Gersten R. (2006, February). *Three-tier mathematics intervention: Emerging model & preliminary findings.* Paper presented at the Pacific Coast Research Conference, San Diego, CA.

The Psychological Corporation. (1995). *Stanford achievement test* (9th ed.). San Antonio, TX: Author.

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

The Psychological Corporation. (2002b). *Wechsler individual achievement test–Second edition (WIAT-II).* San Antonio, TX: Author.

Robinson, R., & Miller, D. (1986). *Kindergarten diagnostic instrument.* Columbus, OH: Kindergarten Interventions and Diagnostic Services.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, 96, 265–282.

Simner, M. L. (1982). Printing errors in kindergarten and the prediction of academic performance. *Journal of Learning Disabilities, 15*, 155–159.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.

Tiesl, J. T., Mazzocco, M. M., & Myers, G. F. (2001). The utility of kindergarten teacher ratings for predicting low academic achievement in first grade. *Journal of Learning Disabilities, 34*, 286–293.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvin, C. (1999). Preventing reading failure in young children with phonological processing difficulties: Group and individual responses to instruction. *Journal of Educational Psychology, 91*, 579–593.

Urbina, S. (2004). *Essentials of psychological testing.* Hoboken, NJ: John Wiley & Sons.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363–382.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137–146.

Wilkinson, G. (1993). *Wide range achievement test-Third edition.* Wilmington, DE: Wide Range.

Woodcock, R. M., & Johnson, M. B. (1989). *Woodcock-Johnson psycho-educational battery-Revised.* Allen, TX: DLM Teaching Resources.

ABOUT THE AUTHORS

PAMELA M. SEETHALER (Tennessee CEC), Research Associate; and **LYNN S. FUCHS** (Tennessee CEC), Joe B. Wyatt Distinguished University Professor, Department of Special Education, Peabody College, Vanderbilt University, Nashville, Tennessee.

Correspondence concerning this article should be addressed to Pamela Seethaler, 228 Peabody College, Vanderbilt University, Nashville, TN 37203 (e-mail: pamela.m.seethaler@vanderbilt.edu).

Manuscript received March 2009; accepted December 2009.