Exceptional Children

Vol. 73, No. 3, pp. 311-330.

Mathematics Screening and Progress Monitoring at First Grade: Implications for Responsiveness to Intervention

LYNN S. FUCHS
DOUGLAS FUCHS
DONALD L. COMPTON
JOAN D. BRYANT
CAROL L. HAMLETT
PAMELA M. SEETHALER
Vanderbilt University

ABSTRACT: *The predictive utility of screening measures for forecasting math disability (MD) at the end of 2nd grade and the predictive and discriminant validity of math progress-monitoring tools were assessed. Participants were 225 students who entered the study in 1st grade and completed data collection at the end of 2nd grade. Screening measures were Number Identification/Counting, Fact Retrieval, Curriculum-Based Measurement (CBM) Computation, and CBM Concepts/Applications. For Number Identification/Counting and CBM Computation, 27 weekly assessments were also collected. MD was defined as below the 10th percentile at the end of 2nd grade on calculations and word problems. Logistic regression showed that the 4-variable screening model produced good and similar fits in accounting for MD–calculation and MD–word problems. Classification accuracy was driven primarily by CBM Concepts/Applications and CBM Computation; CBM Concepts/Applications was the better of these predictors. CBM Computation, but not Number Identification/Counting, demonstrated validity for progress monitoring.*

Individuals with learning disabilities constitute approximately 5% of the school-age population (U.S. Department of Education, 2000). Because of the additional costs involved in educating this population, as well as the potential stigma associated with a disability label, accurate identification is crucial. Yet, the traditional method for identifying these students, which relies on a discrepancy between intelligence and achievement, has come under attack for conceptual and technical problems (see Vaughn & Fuchs, 2003). One model for reorienting learning disabilities identification, codified in the most

recent reauthorization of the federal disabilities law, involves documenting a child's inadequate response to scientifically validated or research-based intervention. The central assumption is that a lack of responsiveness to a generally effective intervention eliminates instructional quality as a viable explanation for poor academic growth and instead provides evidence of a disability.

Most responsiveness-to-intervention (RTI) models of learning disabilities identification occur within a multitier prevention system. Although some RTI systems incorporate four or more layers of intervention, most include three tiers, as follows. General education is the first tier, and students who enter the RTI learning disabilities identification process must first show evidence of failing in this mainstream setting. After that, educators administer a second tier of prevention, which involves scientifically validated or research-based small-group tutoring. Students who show poor response to this second and more intensive tier of intervention enter special education, which is the third and most intensive instructional tier.

In implementing an RTI multitier prevention model for identifying learning disabilities, the first step is determining the students who are at risk for developing learning disabilities and who therefore require attention within the prevention system. Identifying this risk pool early, in kindergarten or first grade, permits students to participate in prevention services before the onset of substantial academic deficits. The goal is to increase the likelihood that the academic competence of these students will develop adequately.

Two types of errors, however, challenge the accuracy of procedures for classifying children into at-risk and not-at-risk groups early in their schooling. The first type of error is false positives, in which children who eventually become academically competent score below the cutpoint on the predictive instrument, so educators therefore consider them to be at risk. False positives undermine prevention by stressing school resources to provide Tier 2 intervention to an inflated percentage of students (Fletcher et al., 2002; Jenkins & O'Connor, 2002). The second error type is false negatives, that is, children who score above the cut score on a predictive instrument but who later develop academic problems. A classification process that produces large numbers of false negatives diminishes the effectiveness of prevention efforts by depriving at-risk children of the early intervention that they require (Jenkins, 2003; Torgesen, 2002). For a multitier prevention system to work effectively, procedures for determining risk must yield a high percentage of true positives while identifying a manageable risk pool by limiting false positives.

The goal is to increase the likelihood that the academic competence of these students will develop adequately.

The first purpose of the present study was to explore methods for screening first-grade children who require attention within a multitier prevention system that incorporates an RTI model for identifying learning disabilities. Our focus was mathematics. The incidence of math disabilities (MD) is similar to the incidence of reading disability (Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994), even though less systematic research has focused on MD. In the next section, we summarize previous work on math screening conducted at the kindergarten and first-grade levels. We then clarify how the present study extends the screening literature and explain the study's second purpose concerning progress monitoring.

PREVIOUS MATH SCREENING STUDIES

We identified nine studies that (a) were conducted at the kindergarten or first-grade level, (b) reported predictive validity information on math screening measures, and (c) incorporated math outcomes. For example, we eliminated Magliocca, Rinaldi, and Stephens (1979) because neither the outcome ("teacher judgment that the student will experience considerable difficulty or be retained") nor the predictors (writing Xs in grids, copying figures, and naming colors) were specific to math. In addition, we did not include Gersten, Jordan, and Flojo (2005) because it is a synthesis paper and includes the same study reported in Baker et al. (2002).

Table 1 summarizes key information for each study. The table displays sample size, grade where screening occurred, grade where outcome was assessed, the screening instruments, the outcome measures, the correlations between the screeners with the various outcomes (identified as outcome 01 or outcome 02 within each study), and the decision utility information (hit rate, or the percentage of cases correctly classified as at high risk for MD and at low risk for MD; sensitivity, or the degree to which a screening measure correctly identifies children at high risk for MD; and specificity, or the degree to which a screening measure correctly identifies children at low risk for MD). Using Baker et al. (2002) as an example, the table reads as follows. The study initially assessed students (n = 64–65) in the spring (S) of kindergarten on four screening measures: Number Knowledge Test (Okamato & Case, 1996), digit span backward, numbers from dictation, and magnitude comparison. Assessment in the spring of first grade involved two outcomes: the Stanford Achievement Test 9 (Harcourt Brace Educational Measurement, 1996) and the Number Knowledge Test. The correlations with respect to outcome 01 (Stanford Achievement Test 9) ranged between .47 and .72. For outcome 02 (Number Knowledge Test), correlations were similar (.45 to .72). Baker et al. provided no information on decision utility to determine the accuracy of the individual predictions of risk with respect to outcomes. Five of the nine studies incorporated, but were not limited to, similar measures (developed at the University of Oregon), most frequently including number identification, quantity discrimination or magnitude comparisons, and missing numbers.

Six studies focused on kindergarten screening. Baker et al. (2002) reported predictive validity from the end of kindergarten to the end of first grade. The coefficients for the multiskill Number Knowledge Test, when used as a screener, appeared higher (.72) than those for the single-skill and perhaps easier screeners, where correlations for numbers from dictation and magnitude comparisons were .45 to .60 (the researchers did not test differences between correlations). Two other kindergarten studies incorporated a similar set of single-skill screeners. Lembke and Foegen (2006), who considered the same predictive time frame, documented coefficients of similar magnitude. By contrast, the predictive interval of Bryant, Bryant, Kim, and Gersten (2006)—which was limited to a time frame of several months—found somewhat higher correlations (including a coefficient of .73 for the total score across five of their measures, excluding backward digit span).

Simner (1982); Teisl, Mazzocco, and Myers (2001); and VanDerHeyden, Witt, Naquin, and Noell (2001)—in contrast to Baker et al. (2002), Lembke and Foegen (2006), and Bryant et al. (2006)—reported or provided information for calculating hit rate, sensitivity, and specificity, sometimes with correlation coefficients (Simner; Teisl et al.) and sometimes without (VanDerHeyden et al.). Teisl et al. was the only one of these three studies to directly assess math outcomes, either 4 months later (spring of kindergarten to fall of first grade) or 1 year later (spring of kindergarten to spring of first grade). With teacher ratings as the screener, the correlation with the Test of Early Mathematics Ability-2 (Ginsburg & Baroody, 1990) was .34, with better specificity (87.7) than sensitivity (65.2). Simner was one of few studies that used a nonmath skill, reflecting a math-related cognitive ability (writing reversals of numbers and letters from short-term memory), as the predictor to screen kindergarten students for future math difficulties. The outcome was kindergarten teachers' rankings of first-grade readiness, where correlations were high (.67) but where specificity (86.7) again exceeded sensitivity (77.3). The correlation for the later outcome, report card grades during the winter of first grade appeared lower (.40), with no predictive utility information reported. Finally, VanDerHeyden et al. examined three kindergarten screeners (the researchers did not indicate time of year) with respect to professional judgments of student difficulties (end-of-kindergarten retention or referral to the school's problem-solving team or identification as a validation problem). For each screener, they again found lower sensitivity (0.00–71.4) than specificity (90.9–94.4), with two of the screeners picking up none of the referrals or validation problems.

We identified three studies conducted at the first-grade level (including Bryant et al., 2006, which was also described at the kindergarten level). Daly, Wright, Kelley, and Martens (1997)

TABLE 1

Summary of Predictive Validity Screening Studies

Author	n	*Grade Screen*	*Grade Outcome*	*Screeners*	*Outcome*	r *01*	r *02*	*Decision Utility HR*	*Decision Utility Sensitivity*	*Decision Utility Specificity*
Baker et al. (2002)	64–65	K(S)	1(S)	Number Knowledge Test (NKT)	SAT-9(01)	.72	.72			
				Digit span backward	NKT(02)	.47	.60			
				Numbers from dictation		.47	.48			
				Magnitude comparison		.54	.45			
Bryant et al. (2006)	104	1(W)	1(S)	Oral counting	SAT-10(01)	.42				
				Number identification		.56				
				Quantity discrimination		.62				
				Missing number		.56				
				Digit span backward		.49				
				Addition facts		.66				
				Total score		.70				
	135	K(W)	K(S)	Oral counting		.49				
				Number identification		.51				
				Quantity discrimination		.61				
				Missing number		.67				
				Digit span backward		.54				
				Total score		.73				
Chard et al. (2005)	168	K(F)	K(S)	Count 10s	NKT(01)	.55				
	207	1(F)	1(S)	Count 5s		.53				
				Count 2s		.49				
				Number writing		.57				
				Number identification		.58				
				Quantity discrimination		.50				
				Missing number		.64				

continues

TABLE 1 *(Continued)*

Author	n	*Grade Screen*	*Grade Outcome*	*Screeners*	*Outcome*	r *01*	r *02*	*Decision Utility* *HR*	*Sensitivity*	*Specificity*
Clarke & Shinn (2004)	52	1(F)	1(S)	Oral counting Number identification Quantity discrimination Missing number	WJ Applied Problems (01) CBM Computation (02)	.72 .72 .79 .72	.56 .60 .70 .67			
Daly et al. (1997)	30	1(F)	(4 mos later)	Number reading Number counting Number production Number selection	Basic facts + (01) Basic facts – (02)	.07 .39 .36 .30	.07 .39 .36 .30			
Lembke & Foegen (2006)	16–20	K(F)	1(F)	Quantity discrimination Number identification Missing number	Stanford Early Achievement (01)	.57 .51 .49				
Simner (1982)	67 53	K(F) K(W)	(S for 01) (W for 02)	Writing reversible numbers and letters from STM (form errors)	Teacher rankings for G1 readiness (01) G1 Report cards (02)	–.67 –.40		83.6%	77.3%	86.7%
Teisl et al. (2001)	234	K(S)	1(F-S)	Teacher ratings of math perf (<10th)	TEMA-2 (01)	0.34		85.4%	65.2%	87.7%
VanDerHeyden et al. (2001)	25	K(?)	K(S)	Circle number Write number Draw circles	Retention Referral to school team "Validation problem"			88.0% 80.0% 88.0%	71.4% 0.0% 0.0%	94.4% 90.9% 91.7%

tested 30 children in the fall on number knowledge, using four formats (number reading, number counting, number production, and number selection) and then again 4 months later on the outcomes: addition basic facts and subtraction basic facts. The correlations ranged between .07 and .39, with identical coefficients for addition and subtraction outcomes. By contrast, Clarke and Shinn (2004) used a greater variety of single-skill screeners and obtained more encouraging results. With 52 first graders, the correlations between the fall screeners (oral counting, number identification, quantity discrimination, and missing number) and the spring outcomes (Woodcock Johnson Applied Problems and curriculum-based measurement computation) ranged from .56 to .79. The .79 occurred for quantity discrimination, and the pattern of coefficients suggested that the screeners were better at predicting concepts and applications than at predicting computation performance (the researchers did not test differences between correlations). Bryant et al. incorporated a similar battery of measures but also included addition facts and reported a coefficient for the total score across the individual tests. The coefficients of Bryant et al. for the single-skill scores ranged from .42 for oral counting to .66 for addition facts, with .70 for the total score across measures. This outcome suggests differential validity for the most difficult screener, addition facts; but the study did not assess differences between correlations, and the prediction involved a shorter time frame, winter to spring of first grade.

The final study (Chard et al., 2005) included a mix of kindergarteners and first graders from fall to spring of the same year and used the Number Knowledge Test as the outcome. These researchers examined 10 screeners. We excluded counting to 20, counting from 3, and counting from 6 because correlations ranged between .07 and .18. The others, shown in Table 1, were skip counting by 10s, 5s, or 2s; number writing; number identification; quantity discrimination; and missing number. Correlations ranged between .49 and .64, with the highest coefficient for missing number.

Across these studies, we offer four observations. First, at first grade, correlations appeared stronger when researchers used the Number Knowledge Test as a screener (Baker et al., 2002; Chard et al., 2005) and when Bryant et al. (2006) considered a total score across six single-skill screeners, including the more difficult addition facts. A key distinguishing feature of the Number Knowledge Test and Bryant et al.'s total score is the assessment of a broader scope of skills, including more difficult items. Of course, both are relatively time-consuming as screeners; the Number Knowledge Test requires more than just a few minutes of individual administration, and Bryant et al.'s total score involves administration of six 1-min measures. Yet, the relatively high coefficients suggest the need for additional study of time-efficient screeners that incorporate a more difficult or broader span of skills. Such an approach seems conceptually sound because previous work (e.g., Fuchs, Fuchs, Stuebing, et al., 2006; Fuchs, Fuchs, Compton, Powell, et al., 2006) indicates that the various aspects of mathematics (e.g., problem solving vs. computation) may involve separate abilities. This outcome challenges the notion that a single math behavior may serve as a viable predictor of mathematics difficulties later in school, as the range of skills required for success grows.

Our second observation is that screening may prove more difficult for detecting students who will emerge as having math difficulties (sensitivity) than for predicting who will develop adequately (specificity). This result may be inevitable because most children will in fact not develop math difficulties (creating a bigger denominator in the calculation of specificity). At the same time, another potential deterrent to sensitivity, related to our third observation, is that the screening measures investigated to date may be insufficiently difficult to yield fine discriminations at the lower end of the distribution. This observation may be especially valid at first grade, where the screening tasks, with the exception of Bryant et al.'s (2006) addition facts, do not extend much beyond the difficulty of the measures studied at kindergarten. Finally, with respect to methodology, none of the existing studies examined outcomes more than 12 months out or beyond first grade, with few studies reporting correlational data as well as decision utility data.

HOW THE PRESENT STUDY EXTENDS PREVIOUS WORK

In the present study, we sought to extend previous work on math screening. Our first substantive extension to the literature concerns the nature and breadth of the screeners examined. We included (a) a relatively easy screener that incorporated a limited set of skills (number identification/counting, e.g., 3, 4, 5, __, __); (b) a more difficult single-skill screener that relied on fact retrieval (e.g., 4 + 2, 9 – 3); (c) a multiple-skill computation screener that sampled the entire first-grade curriculum; and (d) a multiple-skill concepts and applications screener that sampled the entire first-grade curriculum.

The second substantive contribution involves the nature of the predicted outcomes, where we separated computation from word problems performance. Previous work demonstrates that these skills may represent separate dimensions of competence. For example, Fuchs, Fuchs, Stuebing, et al. (2006) assessed 924 third graders on three measures of calculation and four measures of math problem-solving skill. The researchers identified students without difficulty, with calculations difficulty, with word problems difficulty, or with calculations and word problems difficulty. Profile analysis showed distinctive patterns of cognitive abilities associated with calculations versus word problems difficulty, regardless of whether the difficulty occurred alone or in combination. Although calculations and problem solving may comprise two distinct aspects of mathematics cognition, previous screening studies have not explored these outcomes separately. Also with respect to outcomes, we assessed MD at the end of second grade, nearly 2 years after collecting the initial screening data. At most, prior work has explored outcomes 1 year out and at the end of first grade.

In addition to these substantive contributions, we extended previous work in two methodological ways. First, we used logistic regression to look beyond predictive correlations to consider hit rate, sensitivity, and specificity. This focus on decision utility permitted us to assess the accuracy of individual designations of risk. Second, with respect to the designation of MD, we employed a relatively stringent research criterion: below the 10th percentile. The MD literature sometimes employs cutpoints as high as the 35th percentile to designate math difficulty. We, by contrast, selected the 10th percentile so we could generalize findings to school contexts, which reserve the designation of disability for extremely poor performance.

Beyond these substantive and methodological extensions to the math screening literature, we adopted a second purpose: to compare the tenability of two math assessments for the purpose of monitoring students' progress during first grade. As with screening, progress monitoring is a central assessment component within RTI, because progress monitoring can help determine whether students are responding to intervention. For this reason, educators need information about whether progress-monitoring tools validly index the development of mathematics outcomes.

Most previous research on math progress monitoring at first grade, however, has only investigated technical features of performance on progress-monitoring tools at one point in time (see Table 1). Although the reliability and validity of a one-time administration are important in substantiating the tenability of a progress-monitoring tool, determining the validity of the slope is critical to permit conclusions about whether increasing scores, based on frequent measurements over time, are associated with long-term overall competence. Using relatively easy single-skill math screeners at first grade, Clarke (2005) described rates of increase (difference scores) from fall to winter and from winter to spring testing to assess how the rate of increase during the fall compared to the rate of improvement in the spring. Lembke and Foegen (2006), by contrast, measured students monthly from November to January and reported weekly rates of increase that were based on ordinary least-squares regression among 30 first graders. They found small average rates of improvement: 0 to .08 for quantity discrimination, –.16 to .06 for number identification, and –.06 to .04 for missing number. We extended these prior studies by conducting weekly assessments over the course of first grade, as would occur if the measures were used for progress monitoring, on two contrasting types of progress-monitoring measures: a single-skill and a multiskill measure. We examined predictive

validity (i.e., the correlation between the first-grade slope of improvement and end-of-second-grade math performance) and discriminant validity (i.e., differences in slope as a function of end-of-second-grade MD status).

METHOD

Participants

We derived the sample for the present research report from an intervention study conducted with students in 41 first-grade classrooms in six Title I schools and four non-Title I schools in a southeastern metropolitan school district (Fuchs et al., 2005). For the intervention study, we had administered a broad set of math measures in September of first grade to the 667 students for whom we had received parental consent (i.e., 89% of the students in the 41 classrooms). We had identified 139 students with low math performance (i.e., below the 21st percentile) at study entry on a broad array of math measures. We then randomly assigned the low-study-entry students to receive Tier 2 math tutoring (n = 70) or not receive such tutoring (n = 69). For comparison purposes, we also followed an additional 180 students, deemed average at study entry on this broad array of measures, across first and second grades. Because some students moved to other schools, the size of these three groups (low-study-entry tutored, low-study-entry control, and average-study-entry comparisons) had decreased from September to May to 63, 64, and 145, respectively. By the end of second grade, when students had dispersed to 116 classrooms, complete data on the variables of the present study were available for 55, 57, and 113 students in the three respective groups. These 225 students comprised the sample for the progress-monitoring questions posed in the present study. For these 225 students, the average standard scores on the Woodcock-Johnson III (WJ III; Woodcock, McGrew, & Mather, 2001) Word Identification, Calculation, and Applied Problems subtests, respectively, were 105.90 (SD = 11.31), 102.18 (SD = 13.69), and 97.92 (SD = 11.97). The sample included 122 males (54.2%). In the sample, 114 (54.0%) received subsidized lunch. The race/ethnicity distribution was 103 (45.8%) African American, 103 (45.8%) non-Hispanic White, 15 (16.7%) Hispanic, and 4 (1.8%) other. The sample included no English-language learners. A total of 23 students in the sample (10.2%) received some form of special education: Two students had a learning disability, one had attention deficit/hyperactivity disorder, and 20 had speech and language disorders.

As shown in Fuchs et al. (2005), the low-study-entry tutored students scored significantly better than the low-study-entry control students at the end of first grade. Moreover, at the end of second grade, low math status (i.e., below the 10th percentile on the Wide Range Achievement Test–Arithmetic; WRAT–Arithmetic; Wilkinson, 1993) was significantly more prevalent for the low-study-entry control students than for the low-study-entry tutored students. Consequently, the tutoring the low-study-entry tutored students received influenced their outcomes, thereby negating viable predictions that were based on their initial first-grade performance. For this reason, we eliminated the low-study-entry tutored students when conducting our screening analyses, leaving an untutored sample of 170 students. Among these 170 students, the average standard score on the WJ III (Woodcock et al., 2001) Word Identification, Calculation, and Applied Problems subtests were 107.49 (SD = 10.53), 104.26 (SD = 13.11), and 99.55 (SD = 12.18), respectively. The sample included 89 males (52.3%). In the sample, 82 students (48.2%) received subsidized lunch. The race/ethnicity distribution of the sample was 75 (44.1%) African American, 78 (45.9%) non-Hispanic White, 13 (7.6%) Hispanic, and 4 (2.4%) other. A total of 17 students (10%) received some form of special education: One student had a learning disability, and 16 had speech and language disorders.

First-Grade Screening Measures

We administered four screening measures in September of first grade: fact retrieval; CBM Computation (Fuchs, Hamlett, & Fuchs, 1990); Number Identification/Counting (Fuchs & Hamlett, 2005); and CBM Concepts/Applications (Fuchs et al., 1990). For each measure, students received problems on paper and wrote their responses on paper.

Fact Retrieval. To assess competence with fact retrieval, we used Addition and Subtraction Fact Fluency (Fuchs, Hamlett, & Powell, 2003). Addition Fact Fluency comprises 25 addition fact problems with answers from 0 to 12, presented horizontally on one page. Subtraction Fact Fluency comprises 25 subtraction fact problems with answers from 0 to 12, presented horizontally on one page. Each of these tests allows students 1 min to write answers. The score is the number of correct answers across the addition and subtraction subtests. Staff entered responses into a computerized scoring program on an item-by-item basis, and an independent scorer reentered 15% of the tests. Data-entry agreement was 99.4%. Coefficient alpha was .84.

CBM Computation. CBM Computation (Fuchs et al., 1990) is a one-page test displaying 25 items that sample the typical first-grade computation curriculum: Problems with two single-digit number requiring adding or subtracting, problems with three single-digit number requiring adding, and problems with two double-digit number requiring adding or subtracting without regrouping. Students have 2 min to complete as many problems as possible. The score is the number of digits correct. Staff entered responses into a computerized scoring program on an item-by-item basis, and an independent scorer reentered 15% of the tests. Data-entry agreement was 99.6. Coefficient alpha was .96.

Number Identification/Counting. Number Identification/Counting (Fuchs & Hamlett, 2005) is a four-item test that presents students with number sequences, the last two of which are shown as blanks (e.g., 4, 5, 6, ___ , ___). The student writes numerals to complete the sequence. Students have 1 min to complete the items, and the score is the number of items completed with two correct numbers, aggregated across two tests taken 1 week apart, for a maximum score of 8. Staff entered responses into a computerized scoring program on an item-by-item basis, and an independent scorer reentered 15% of tests. Data-entry agreement was 99.3. Coefficient alpha was .92.

CBM Concepts/Applications. CBM Concepts/Applications (Fuchs et al., 1990) is a three-page test with 25 items that sample the typical first-grade concepts and applications curriculum, including numeration, concepts, geometry, measurement, applied computation, charts and graphs, and word problems. The tester reads the words in each item aloud. For 20 items, students have 15 s to respond before the tester reads the next item; for the remaining 5 items, students have 30 s. The score is the number of correct answers. Staff entered responses into a computerized scoring program on an item-by-item basis, and an independent scorer reentered 15% of tests. Data-entry agreement was 97.8%. Coefficient alpha was .92.

First-Grade Progress-Monitoring Measures

The researchers administered two measures, CBM Computation and Number Identification/Counting, weekly in large-group format. For CBM Computation, each weekly alternate form (a) samples the annual computation curriculum with the same problem types in the same proportions and (b) displays problems in random order. Students have 2 min to write answers, and the score is the number of digits correct. The maximum score is 31 or 32, depending on the form (on one alternate form, the maximum is 34, because sums of 10 are randomly drawn more frequently). Staff entered responses on an item-by-item basis into software that scored performance. A second scorer independently reentered 15% of the protocols, with 98.5% agreement. Performance consists of weekly rate of improvement (calculated with an ordinary least-squares regression between score and calendar day). Coefficient alpha (estimated at Week 10) was .90.

For Number Identification/Counting, each weekly alternate form samples four items, and students have 1 min to write their responses to these items. The score is the number of items completed with two correct numbers, aggregated across 2 consecutive weeks, for a maximum score of 8. Staff entered responses on an item-by-item basis into software that scored performance. A second scorer independently reentered 15% of the protocols, with 99.3% agreement. Performance consisted of the weekly rate of improvement (calculated with an ordinary least-squares regression between score and calendar day). Coefficient alpha (estimated at Week 10) was .89.

Second-Grade Outcome Measures and the Designation of MD

In the spring of second grade, the researchers administered two outcome measures: the WRAT 3–Arithmetic (Wilkinson, 1993) and Jordan's Story Problems (Jordan & Hanich, 2000).

WRAT 3–Arithmetic (WRAT 3). The WRAT 3–Arithmetic (Wilkinson, 1993) allows students 10 min to complete calculation problems of increasing difficulty. If the basal is not met, students read numerals aloud to the examiner. Median reliability is .94 for ages 5 to 12 years. The correlation with the WJ III Calculation Test at Grade 2 was .71.

Word Problems. Following Jordan and Hanich (2000; adapted from Carpenter & Moser, 1984; Riley & Greeno, 1988; Riley, Greeno, & Heller, 1983), Story Problems comprises 14 brief story problems involving sums or minuends of 9 or less, with change, combine, compare, and equalize relationships. The tester reads each item aloud; students have 30 s to respond. The score is the number of correct answers. A second scorer independently rescored 15% of protocols, with agreement of 99.8%. Coefficient alpha on this sample was .87. We used this measure to index word problems disability because no other measure provides adequate behavior sampling of second-grade word problems.

Designation of MD. Students received a designation as MD in two ways: (a) scoring below the 10th percentile on the WRAT 3–Arithmetic (i.e., MD–calculation) using the WRAT 3 national norms, and (b) scoring below the 10th percentile on Story Problems (i.e., MD–word problems). We derived the normative profile for designating disability on Story Problems from a local but representative sample of 634 students at the end of second grade, external to the present study.

Procedure

Administration of Screening and Outcome Measures. Before administering each measure, staff learned, practiced, and established agreement on administration procedures. For each measure, examiners, who were research assistants on this project, used a standard script from which they read directions. Examiners administered the screening measures in large groups. Two research assistants and the classroom teacher supervised each session. Examiners administered the outcome measures individually or in small groups. Students who were absent for the screening completed make-up testing individually or in small groups.

Administration and Use of Weekly Progress-Monitoring Measures. Research assistants administered the first weekly CBM Computation and Number Identification/Counting tests to intact classes while teachers observed. We then gave teachers administration scripts and audiotapes with start and stop times. Each week, staff provided teachers with a new set of tests that included students' names; teachers administered the tests in whole-class format; and research assistants picked up the week's completed tests. Research assistants entered students' responses into software for automatic scoring and data management. On CBM Computation and Number Identification/Counting, data-entry agreement calculated on 20% of protocols was 99.4% and 99.8%, respectively, at Weeks 3, 10, and 15. After seven weekly administrations and every 2 weeks thereafter, research assistants printed computerized reports and met with teachers to discuss the reports, which summarized the class performance for the most recent 2-week interval. At Week 10, to teach students how to interpret their graphs, research assistants printed computerized student graphs and used a scripted lesson with overheads to conduct a whole-class session. Every 3 weeks thereafter, research assistants provided teachers with graphs to share with their students. Although teacher and student feedback may possibly affect student trajectories, we opted to provide such feedback to mirror typical progress monitoring in schools.

Data Analysis

Screening. We used logistic regression to predict membership in the second-grade MD and non-MD (NMD) groups, separately for MD–calculation and for MD–word problems. Binary logistic regression is a form of regression used to predict a dichotomous dependent variable on the basis of independent variables. It provides the percentage of variance in the dependent variable explained by the independents and ranks the

relative importance of independent variables. The output from logistic regression is a set of coefficients for an equation that calculates the probability that a new case is of a certain class. Logistic regression is relatively flexible because it does not assume (a) a linear relation between dependents and independents, (b) a normally distributed dependent variable, (c) homogeneity of variance, (d) normally distributed error terms, (e) interval-level independents, or (f) unbounded independents. Logistic regression assumes that multicollinearity does not occur and that outliers are not a problem; we tested these assumptions, which proved viable.

For the logistic regression models in this study, we were most interested in maximizing the rate of true positives and then observing the associated rate of false positives. In the context of RTI, the sum of true and false positives constitutes the sample for Tier 2 intervention. For this reason, we specified the classification cutoff for the logistic regression models to be equal to the proportion of second-grade MD children in the sample (i.e., these are the children that the first-grade screeners should have identified). We generated logistic regression models by using SPSS 13.0 statistical software, and we entered all four predictors together in one block.

To contrast various classification models, we used sensitivity, specificity, and area under the ROC curves (generated by using SPSS 13.0). The calculation for *sensitivity*, or the degree to which a screening measure correctly identifies children at high risk for MD (i.e., true positives), involves dividing the number of true positives by the sum of true positives and false negatives. Sensitivity increases as the number of false negatives decreases. *Specificity*, on the other hand, refers to the degree to which a screening measure correctly identifies children at low risk for MD (i.e., true negatives) and is calculated by dividing the number of true negatives by the sum of true negatives and false positives. Specificity increases as the number of false positives decreases. An *ROC curve* is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different possible cutpoints of a diagnostic test.

For evaluating differences in predictive accuracy across models (see Steadman et al., 2000; Tsien, Fraser, Long, & Kennedy, 1998), we used area under the ROC curve (AUC), a measure of discrimination, or the ability of the screening test to correctly classify second-grade MD and NMD children. For example, if we had already correctly classified children into MD and NMD groups and then randomly picked and tested one child from the MD group and one from the NMD group, we would expect that the child scoring lower on the first-grade prediction battery would be the one from the MD group. The AUC, which represents the percentage of randomly drawn pairs for which the screening test correctly classifies the two children in the random pair, ranges from .5 (i.e., chance performance) to 1.0 (i.e., perfect performance; Swets, 1992). We considered an AUC greater than .90 excellent; .80 to .90, good; .70 to .80, fair; and below .70, poor. A lack of overlap on the confidence intervals on the AUCs for two models indicated differences in predictive accuracy across the models.

Progress Monitoring. To examine how CBM Computation and Number Identification/Counting functioned as progress-monitoring tools, we assessed predictive validity by calculating the correlation between slope and end-of-second-grade performance on the WRAT 3–Arithmetic and on Jordan's Story Problems. We tested the difference between correlations with WRAT and Story Problems by using the formula of Walker and Lev (1953) for testing differences between correlations calculated on dependent samples. To assess discriminant validity, we conducted one-way analyses of variance (ANOVAs) on slopes, using end-of-second-grade MD status as the between-subjects variable, and also reported corresponding effect sizes (i.e., the difference between the NMD and MD means, divided by the standard deviation of the NMD group).

RESULTS

MD Prevalence

Of the 170 students in the sample, we designated 23 as MD–calculation and 24 as MD–word problems. Eight qualified as MD–calculation only, 8 were MD–word problems only, and 31 were both. To estimate prevalence rates for the general population, we extrapolated these figures to the population of 667 first-grade consented students

TABLE 2

Demographics by Second-Grade MD Status

	Calculation					*Word Problems*				
	NMD (n = 147)		*MD (n = 23)*			*NMD (n = 146)*		*MD (n = 24)*		
Variable	n	*(%)*	n	*(%)*	χ^2	n	*(%)*	n	*(%)*	χ^2
Males	73	(49.7)	16	(69.6)	(1)3.16	73	(50.0)	16	(67.7)	(1)2.29
Race: African American	65	(44.2)	10	(43.5)	(3)3.32	64	(43.8)	11	(45.8)	(3)1.22
Caucasian	65	(44.2)	13	(56.5)		66	(45.2)	12	(50.0)	
Hispanic	13	(8.8)	0	(0.0)		12	(8.2)	1	(4.2)	
Other	4	(2.7)	0	(0.0)		4	(2.7)	0	(0.0)	
Subsidized Lunch	72	(49.0)	10	(43.5)	(1)0.02	69	(47.2)	13	(54.2)	(1)0.58

Note: NMD = not math disabled; MD = math disabled.

as follows. We first used the percentage of MD among the low-study-entry control students represented in the present sample (20 of 57, or 35.1%, for MD–calculation and 19 of 57, or 33.3%, for MD–word problems) to estimate the number of students with MD among the low-study-entry students not represented in the present sample (because they were part of the low-study-entry tutored group [n = 70] or because they were low-study-control students with missing data [n = 12]). We thereby estimated that without intervention, we would have identified 29 of these 82 low-study-entry students with MD–calculation and 27 with MD–word problems. We next used the percentage of MD among the average-study-entry comparison students represented in the present sample (3 of 113, or 2.7%, for MD–calculation and 5 of 113, or 4.4%, for MD–word problems) to estimate the number of students with MD among the average-study-entry comparison students not represented in the present sample (because they were average-study-entry students not sampled for comparison purposes [n = 348] or because they were average-study-entry comparison students with missing data [n = 35]). We thereby estimated that we would have identified 10 of these 383 students with MD–calculation and that we would have identified 17 of them with MD–word problems. This method resulted in projecting that a total of 62 of the 667 students would have MD–calculation, for a prevalence rate of 9.3%, and that 68 students would have MD–word problems, for a prevalence rate of 10.2%.

Although we set our cutpoint for MD at the 10th percentile, we relied on normative frameworks that were external to the 41 classrooms from which we drew our sample. We therefore did not necessarily expect to net 10% of the students in these 41 classrooms as MD. Moreover, because we did not sample all students in these 41 classrooms (we omitted the higher-performing students), we believe that we missed at least a few students who would have emerged as MD at the end of second grade, rendering our figures of 9.3% and 10.2% underestimates of MD prevalence.

In Table 2, we show sex, race, and subsidized lunch status by MD–calculation status and by MD–word problems status, as well as chi-square values testing whether these demographic variables related to MD status. Table 3 shows means and standard deviations by MD–calculation status and by MD–word problems status on (a) beginning-of-first-grade performance on the WJ III subtests, (b) beginning-of-first-grade performance on the four screening measures, and (c) end-of-second-grade performance on WRAT–Arithmetic and Story Problems. For each variable, we also report results of a one-way analysis of variance (NMD vs. MD) with F value and effect size. Interestingly, as shown, the MD and NMD groups were demographically comparable. By contrast, as would be expected, beginning-of-first-grade and

TABLE 3

Performance at Beginning of Grade 1, Performance at End of Grade 2, and Slope During Grade 1

	Calculation						*Word Problems*					
	NMD (n = 147)		*MD (n = 23)*				*NMD (n = 146)*		*MD (n = 24)*			
Variable	M	(SD)	M	(SD)	F	*ES*	M	(SD)	M	(SD)	F	*ES*
Beginning Grade 1: Descriptive												
WJ-WID	108.80	(9.19)	99.39	(14.34)	17.39***	1.02	108.85	(9.31)	99.46	(13.58)	18.04***	1.01
WJ-Cal.	106.22	(12.28)	91.74	(11.36)	27.47***	1.18	106.38	(12.40)	91.42	(9.62)	30.94***	1.21
WJ-AP	100.93	(12.06)	90.70	(8.90)	15.13***	0.85	100.79	(12.36)	91.96	(7.52)	11.45**	0.71
Beginning Grade 1: Screeners												
Fact Retrieval	1.51	(1.60)	0.91	(1.04)	3.01	0.38	1.52	(1.60)	0.88	(1.03)	3.65	0.40
CBM Comp.	7.99	(3.73)	4.22	(3.19)	19.19***	1.01	8.00	(3.92)	4.33	(3.40)	18.70***	0.94
Number ID/Count.	5.42	(2.46)	3.04	(2.55)	18.20***	0.96	5.34	(2.44)	3.34	(2.93)	13.76***	0.82
CBM Concepts/App.	10.00	(2.59)	7.00	(2.68)	26.47***	1.16	9.99	(2.58)	7.17	(2.87)	24.05***	1.09
End Grade 2: Outcomes												
WRAT-Arith.	99.61	(8.26)	71.04	(8.41)	236.61***	3.46	99.00	(9.29)	75.92	(13.58)	110.08***	2.48
Word Problems	95.05	(14.82)	69.35	(8.38)	65.65***	1.73	95.80	(13.90)	65.89	(3.51)	109.45***	2.15
First-grade slope												
CBM Comp.	0.37	(0.29)	0.22	(0.17)	6.09*	0.52	0.37	(0.29)	0.24	(0.18)	4.60*	0.45
Number ID/Count.	0.07	(0.08)	0.09	(0.06)	1.15	–0.25	0.07	(0.08)	0.09	(0.07)	1.84	–0.25

Note. NMD = not math disabled; MD = math disabled. CBM Comp = Curriculum-Based Measurement Computation; Number ID/Count = Number Identification/Counting; CBM Concepts/App. = Curriculum-Based Measurement Concepts/Applications; WRAT-Arith. = Wide-Range Achievement Test 3-Arithmetic; Word Problems = Jordan's Story Problems.

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 4

Correlations Between Beginning First-Grade Screeners and End-of-Second-Grade Outcomes (n = *170*)

	Variables				
Variables	*FR*	*CBM-C*	*NIDC*	*CBM-C/A*	*Arith*
Initial Fact Retrieval (FR)					
CBM Comp. (CBM-C)	.336**				
Number ID/Count (NIDC)	.109	.234*			
CBM Concepts/App. (CBM-C/A)	.153	.350**	.360**		
End 2nd-Grade WRAT-Arith. (Arith)	.135	.344**	.340**	.403**	
Word Problems (WP)	.095	.346**	.390**	.442**	.607**

Note. CBM Comp = Curriculum-Based Measurement Computation; Number ID/Count = Number Identification/Counting; CBM Concepts/App. = Curriculum-Based Measurement Concepts/Applications; WRAT-Arith = Wide-Range Achievement Test 3-Arithmetic; Word Problems = Jordan's Story Problems.
*$p < .01$. **$p < .001$.

end-of-second-grade academic performance tended to differ as a function of MD status on both reading and math variables. The only performance variable on which students did not differ was fact retrieval, probably because of a floor effect on the measure.

Screening

In Table 4, we report zero-order correlations between the four screening measures and the two outcome measures. In Table 5, we report the results of the logistic regression analyses for determining how the four first-grade screeners contributed to the prediction of second-grade MD status, along with hit rate, sensitivity, specificity, and area under the ROC curve. In the same table, we present these statistics for the prediction of second-grade MD–word problems status.

For predicting MD–calculation status, the four-variable screening model resulted in a hit rate of 78.2%, with specificity exceeding sensitivity. The AUC was .847, which is deemed good. Among the screening errors, the initial screening missed 7 second graders with MD–calculation, whereas 30 students whom screening identified as at risk did not meet the second-grade criteria for MD–calculation. Among the four screeners, CBM Concepts/Applications contributed most to the prediction of second-grade MD, followed by CBM Computation. Number Identification/Counting and fact retrieval failed to contribute significantly to the prediction when the other two variables were in the mix.

In predicting MD–word problems status, the four-variable screening model resulted in a hit rate of 74.7%, with more similar specificity and sensitivity. The AUC was again good, with a coefficient of .806. Among the screening errors, the initial screening missed 7 second graders with MD–word problems, whereas 36 students whom screening identified as at risk for difficulty did not meet the second-grade criteria for MD–word problems. Among the four screeners, CBM Concepts/Applications and CBM Computation contributed significantly to the prediction of second-grade MD–word problems; Number Identification/Counting and fact retrieval did not.

Progress Monitoring

The correlation of CBM Computation slope with end-of-second-grade WRAT 3–Arithmetic performance was .34, and correlation with end-of-second-grade word problems performance was .28. The correlation of Number Identification/Counting slope with end-of-second-grade WRAT 3–Arithmetic performance was –.11, and correlation with end-of-second-grade word problems performance was –.19. The difference in the correlations (CBM Computation vs. Number Identification/Counting) with end-of-second-grade

TABLE 5

Classification Indices for Logistic Regression Models for MD-Calculation and MD-Word Problems (n = *170)*

											ROC		
Outcome/Model	B	SE	*Wald*	p	*TN*	*FN*	*TP*	*FP*	*Hit Rate*	*Sensitivity*	*Specificity*	*AUC*	SE	*CI*
MD-Calculation														
Number ID/Count	–.384	.197	3.803	.051	117	7	16	30	78.2	69.6	79.6	.847	.043	.763–.932
Fact Retrieval	–.072	.249	0.083	.773										
CBM Comp.	–.199	.087	5.286	.022										
CBM Concepts/App.	–.280	.110	6.495	.011										
Constant	2.633	.876	9.025	.003										
MD-Word Problems														
Number ID/Count	–.283	.190	2.215	.137	110	7	17	36	74.7	70.8	75.3	.806	.053	.703–.910
Fact Retrieval	–.162	.243	0.444	.505										
CBM Comp.	–.193	.083	5.468	.019										
CBM Concepts/App.	–.253	.104	5.928	.015										
Constant	2.364	.839	7.931	.005										

Note. MD = math disabled. CBM Comp = Curriculum-Based Measurement Computation; Number ID/Count = Number Identification/Counting; CBM Concepts/App. = Curriculum-Based Measurement Concepts/Applications; WRAT-Arith = Wide-Range Achievement Test 3-Arithmetic; Word Problems = Jordan's Story Problems. TN = true negatives; FN = false negatives; TP = true positives; and FP = false positives.

WRAT performance was statistically significant, $t(222) = 4.58$, $p < .001$, as was the difference in correlations with end-of-second-grade word-problem performance, $t(222) = 4.60$, $p < .001$.

At the bottom of Table 3, we report first-grade slope on CBM Computation and on Number Identification/Counting; and for each slope, we report results of a one-way analysis of variance (NMD vs. MD) with F value and effect size. As shown in Table 3, CBM Computation slope differed significantly as a function of second-grade MD status for calculation and on word problems, with moderate effect sizes. By contrast, Number Identification/Counting slope did not differ reliably as a function of second-grade MD group status. The negative correlations and the lack of discriminant validity between NMD and MD for Number Identification/Counting probably occurred because large numbers of students reached the ceiling on this measure: At Week 2, 25.3% of the sample had obtained the maximum score; by Week 5, 37.1% had reached it; and by Week 16, 72.0%.

DISCUSSION

We considered the value of a four-variable first-grade screening battery for forecasting calculations and word problems MD at the end of second grade. The four screeners were (a) a relatively easy screener that incorporated a limited set of skills (number identification/counting, e.g., 3, 4, 5, __, __); (b) a more difficult single-skill screener that relied on fact retrieval (e.g., 4 + 2, 9 – 3); (c) a multiple-skill computation screener that sampled the entire first-grade curriculum; and (d) a multiple-skill concepts and applications screener that sampled the entire first-grade curriculum. For specifying risk for MD–calculation and MD–word problems, the four-variable model produced similarly good fits, with an AUC of .847 for MD–calculation and .806 for MD–word problems. In both models, CBM Concepts and Applications and CBM Computation, the multiple-skill screeners that sampled the entire first-grade curriculum, were the primary drivers of classification accuracy. Moreover, CBM Concepts and Applications was the better of these two predictors in specifying both types of MD. In this way, the results echo the findings of Baker et al. (2002). Those findings suggest that multiskill screeners, which tap items across a range of difficulty and math domains, may provide differentially strong correlates of math outcome.

This recurring pattern suggests the potential utility of screeners that incorporate a range of math skills. Why might a multiskill screener, like CBM Concepts/Applications, operate better than the single-skill screeners that we included? First, research (e.g., Fuchs, Fuchs, Stuebing, et al., 2006; Fuchs, Fuchs, Compton, Powell, et al., 2006) indicates that different aspects of mathematics (e.g., problem solving or computation) may represent distinct forms of mathematical cognition. This finding challenges the notion that a single math task may serve as a viable predictor of subsequent mathematics difficulties in school as the range of skills required for success grows. In a related way, as the domain of skills in the screener expands (for example, CBM Concepts/Applications incorporates measurement, concepts, word problems, numeration, applied computation, and more), the screener assesses a larger set of competencies, which may connect in more robust ways with subsequent math performance in school. Yet another reason that a multiskill screener might work better than a single-skill screener is that the relatively easy single-skill screening tasks may yield insufficiently fine discriminations at the lower end of the distribution. This outcome may be especially true at first grade, where the screening behaviors prior to the present study (see Table 1, with the exception of Bryant et al.'s [2006] addition facts) have strongly resembled or have been identical to the screening measures investigated at kindergarten.

For both outcomes, however, the multivariate screening battery did not specify risk completely. For MD–calculation, 30 students who the screening indicated were at risk did not meet the second-grade criterion for MD–calculation, so an RTI-LD identification system that relied on this four-variable screening battery would have tutored 30 students unnecessarily. Moreover, the initial screening missed 7 students who went on to develop MD–calculation by the end of second grade, and they therefore would have been denied the preventive attention that they required. The figures for MD–word problems were similar. A

total of 36 students who the screening indicated were at risk did not meet the second-grade criterion for MD–word problems, so three dozen students would have received tutoring unnecessarily within an RTI-LD identification system. In addition, the initial screening missed another 7 students who emerged with MD–word problems by the end of second grade and who therefore would have been denied the opportunity for Tier 2 tutoring. Within the context of reading, Jenkins and O'Connor (2002) suggested that to function adequately within an RTI system, the screening process needs to select all true positives while limiting false positives. Although such an ambitious standard may be difficult to achieve, educators and researchers need to fine-tune a system for screening MD risk. For example, research might consider including an easier fact retrieval measure, similar to the one that Bryant et al. (2006) used, and might consider conducting short-term progress monitoring for students below the 50th percentile on the screening battery for verifying risk status.

How do our findings compare with previous work on screening? Drawing such comparisons is difficult for several reasons. First, previous studies that report (or provide the basis for deriving) decision utility data—including hit rate, sensitivity, and specificity—have relied on teacher judgments of performance either as screeners or as the criterion for performance; the present study relied instead on students' test performance for the screener as well as an objective and stringent criterion of MD (test performance below the 10th percentile). In addition, the remaining studies relied entirely on correlation coefficients, which are sometimes lower (Daly et al., 1997) and sometimes higher (Baker et al., 2002; Bryant et al., 2006; Chard et al., 2005; Clarke & Shinn, 2004), but we set a higher standard for our criterion for outcome by considering performance at the end of second grade, nearly 2 school years after screening occurred. By contrast, previous work has examined outcomes that are at most 1 year later and no further out than the end of first grade. In addition, in the present study, as opposed to prior work, we restricted participation to students who began first grade as low or average in math (i.e., we excluded high performers). We excluded high performers because we were primarily interested in the accuracy of distinguishing MD from NMD, where distinctions in the lower half of the distribution are most pertinent. Excluding high performers did, however, restrict the range of our correlations and thereby lower them. Moreover, in a similar way, hit rate, specificity, and sensitivity would have been higher if the sample had included a representative distribution that included high performers, for whom forecasting classification would have been easier.

In interpreting this study's screening findings, we note that the relatively easy single-skill screener, Number Identification/Counting, although not a significant contributor, may hold some promise for forecasting second-grade MD, at least for Calculations status. This outcome is reminiscent of prior work, in which number identification, missing number, and counting screeners used at kindergarten and first grade have provided moderate correlations with outcomes several months to 1 year out. Therefore, additional research, employing larger samples and incorporating a greater number of items on the Number Identification/Counting measure (as in prior work), should continue to explore the possibility that number identification and counting may help identify risk for MD, at least in calculations.

Considering this possibility, we can contemplate our study's second purpose: comparing the tenability of two math assessments for the purpose of monitoring student progress across first grade. Progress-monitoring data were available for two measures: CBM Computation (a multiskill task) and Number Identification/Counting (a relatively easy single-skill task reminiscent of those investigated in prior studies). Clarke (2005) had examined the utility of relatively easy single-skill screeners as progress-monitoring tools by comparing rates of increase (difference scores) from fall to winter as opposed to comparing them from winter to spring. Lembke and Foegen (2006) had also measured students on relatively easy single-skill tasks, but did so monthly from November to January. We extended this work by conducting 27 weekly assessments over the course of first grade, as would occur if the measures were in fact used for progress monitoring. We compared the predictive validity (i.e., correlations of the slopes with end-of-second-grade math performance for

the two progress-monitoring measures) and discriminant validity (i.e., contrasting slope as a function of end-of-second-grade MD status) for the relatively easy single-skill task against the broader multiskill task.

The results demonstrated the superiority of the multiskill CBM Computation task, in which the relation between slope and second-grade outcomes was significantly higher than for Number Identification/Counting. At the same time, CBM Computation slopes for NMD students were significantly and substantially higher than for MD–calculation and for MD–word problems students. We did not obtain this result, however, for Number Identification/Counting, where slopes were unsuitably low, a finding that replicates the data presented by Lembke and Foegen (2006).

Moreover, we remind readers that the present study excluded high-performing students. Assessing a representative sample that included these high performers would probably have magnified the findings favoring the CBM Computation slopes. That is, slopes for the CBM Computation task, which did not suffer a ceiling effect, would have been higher and would have discriminated MD status better, whereas slopes for the Number Identification/Counting task, which did suffer a ceiling effect, would have been lower and would not have discriminated MD status as well. Consequently, despite the potential of Number Identification/Counting as a screener (at least for second-grade MD–calculation status and pending additional study with larger sample sizes), it proved problematic as a progress-monitoring tool. This outcome occurred because of a ceiling effect well before the end of first grade, creating a situation in which some students' slopes were indistinguishable from those of low-performing students. This finding, of course, is problematic for gauging responsiveness to instruction to identify students with MD within an RTI framework, and future research should consider a more challenging form of concept/applications progress monitoring, perhaps in the form of CBM Concepts/Applications or some other multiskill measure that incorporates more items and items with greater difficulty than Number Identification/Counting. More generally, the results illustrate how a measure that may provide useful screening data does not necessarily function well to monitor progress.

Clearly, a need exists for more research to delineate optimal methods for screening MD at the beginning of first grade and for monitoring math development across first grade. While the field awaits additional research, however, the present findings can help guide schools that are struggling to identify their assessment methods within a multitier prevention/LD identification system in math. Our results suggest that educators can efficiently use CBM Concepts/Applications and CBM Computation in a first-cut effort to identify risk for poor math development, perhaps in conjunction with short-term progress monitoring to verify risk status. For progress monitoring, study findings indicate that CBM Computation may provide valid information about the development of math competence across first grade.

A need exists for more research to delineate optimal methods for screening MD at the beginning of first grade and for monitoring math development across first grade.

REFERENCES

Baker, S., Gersten, R., Flojo, J., Karz, 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.

Bryant, D. P., Bryant, B. R., Kim, S. A., & Gersten, R. (2006, February). *Three-tier mathematics intervention: Emerging model and preliminary findings.* Poster presented at the 14th annual meeting of the Pacific Coast Research Conference, San Diego, CA.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal of Research in Mathematics Education, 15,* 179–203.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention, 30*(2), 3–14.

Clarke, B. (2005, June). *Validity of screening measures in math at first grade.* Paper presented at a working meet-

ing of the National Center on Learning Disabilities, New York City.

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.

Daly, E. J., Wright, J. A., Kelley, S. Q., & Martens, B. K. (1997). Measures of early academic skills: Reliability and validity with a first grade sample. *School Psychology Quarterly, 12,* 268–280.

Fletcher, J. M., Foorman, B. R., Boudousquie, A., Barnes, M. A., Schatschneider, C., & Francis, D. J. (2002). Assessment of reading and learning disabilities a research-based intervention-oriented approach. *Journal of School Psychology, 40,* 27–63.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97,* 493–513.

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L, & Lambert, W. E. (2006). Problem-solving and computational skill: Are they shared or distinct forms of mathematical cognition? Manuscript submitted for publication.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. C., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology, 98,* 29–43.

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

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Curriculum-based math computation and concepts/applications.* (Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203)

Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). *Fact fluency assessment.* (Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203)

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. P., & Baroody, A. J. (1990). *Test of Early Mathematics Ability* (2nd ed.). Austin, TX: Pro-Ed.

Gross-Tsur, V., Manor, O., & Shalev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology, 37,* 906–914.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test* (9th ed.). Orlando, FL: Harcourt.

Jenkins, J. R. (2003, December). *Candidate measures for screening at-risk students.* Paper presented at the Conference on Response to Intervention as Learning Disabilities Identification, sponsored by the National Research Center on Learning Disabilities, Kansas City, MO.

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: Lawrence Erlbaum.

Jordan, N. C., & Hanich, L. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of Learning Disabilities, 33,* 567–578.

Lembke, E., & Foegen, A. (2006, February). *Monitoring student progress in early math.* Paper presented at the 14th annual meeting of the Pacific Coast Research Conference, San Diego, CA.

Lewis, C., Hitch, G. J., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology and Psychiatry, 35,* 283–292.

Magliocca, L. A., Rinaldi, R. T., & Stephens, T. M. (1979). A field test of a frequency sampling screening instrument for early identification of at risk children: A report on the second year pilot study. *Child Study Journal, 9,* 213–229.

Okamato, Y., & Case, R. (1996). Exploring the microstructure of children's 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.

Riley M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5,* 49–101.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. *The development of mathematical thinking* (pp. 153–196). Orlando: Academic Press.

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

Steadman, H. J., Silver, E., Monahan, J., Applebaum, P. S., Robbins, P. C., Mulvey, E. P., et al. (2000).

A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24,* 83–100.

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

Teisl, 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. (2002). Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 565–613). Mahwah, NJ: Lawrence Erlbaum.

Tsien, C. L., Fraser, H. S. F., Long, W. J., & Kennedy, R. L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infarction. *Medinfo, 9,* 493–497.

U.S. Department of Education. (2000). *Twenty-second annual report to Congress on the implementation of the Individuals with Disabilities Education Act.* Washington, DC: Government Printing Office.

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.

Walker, H. M., & Lev, J. (1953). *Statistical inference.* New York: Holt & Co.

Wilkinson, G. S. (1993). *Wide Range Achievement Test 3.* Wilmington, DE: Wide Range.

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

ABOUT THE AUTHORS

LYNN S. FUCHS (CEC TN Federation), Nicholas Hobbs Professor of Special Education and Human Development; **DOUGLAS FUCHS** (CEC TN Federation), Nicholas Hobbs Professor of Special Education and Human Development; **DONALD L. COMPTON** (CEC TN Federation), Associate Professor; **JOAN D. BRYANT,** Research Associate; **CAROL L. HAMLETT** (CEC TN Federation), Research Associate; and **PAMELA M. SEETHALER** (CEC TN Federation), Doctoral Candidate, Special Education Department), Vanderbilt University, Nashville, Tennessee.

Address correspondence to Lynn S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203 (e-mail: lynn.fuchs@vanderbilt.edu).

This research was supported in part by Grant No. H324U010004 from the U.S. Department of Education, Office of Special Education Programs, and Core Grant No. HD15052 from the National Institute of Child Health and Human Development to Vanderbilt University. Statements do not reflect the position or policy of these agencies, and no official endorsement by them should be inferred.

Manuscript received February 2006; accepted May 2006.