

THE PREDICTIVE UTILITY OF KINDERGARTEN SCREENING FOR MATH
DIFFICULTY: HOW, WHEN, AND WITH RESPECT TO WHAT
OUTCOME SHOULD IT OCCUR?

By

Pamela M. Seethaler

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements for
the degree of

DOCTOR OF PHILOSOPHY

in

Special Education

December, 2008

Nashville, Tennessee

Approved:

Professor Lynn S. Fuchs

Professor Daniel H. Ashmead

Professor Donald L. Compton

Professor Douglas Fuchs

Professor Kimberly J. Paulsen

TABLE OF CONTENTS

	Page
LIST OF TABLES	iii
Chapter	
I. INTRODUCTION	1
II. METHOD.....	14
Participants.....	14
Kindergarten screening measures.....	15
Computation fluency	15
Number sense	18
Quantity discrimination	19
Outcome measures and MD designation	20
Early math diagnostic assessment math reasoning and numerical operations.....	20
Keymath-revised numeration and estimation.....	21
CBM computation and concepts/applications.....	21
MD designation	21
Interscorer agreement	22
Procedure	22
Data analysis	24
Reliability of the screening measures	24
Correlations among screening and outcome measures	24
Logistic regression to predict MD.....	24
ROC curves to contrast various models.....	25
III. RESULTS	27
Descriptive statistics.....	27
Technical adequacy of kindergarten screening measures.....	28
MD prevalence as a function of mathematics outcome.....	36
ROC curves to contrast the predictive utility of logistic regression models	36
IV. DISCUSSION.....	41
Appendix	
A. COMPUTATION FLUENCY MEASURE.....	49
B. NUMBER SENSE MEASURE.....	50
C. NUMBER SENSE MEASURE SCORE SHEET	55
REFERENCES	56

LIST OF TABLES

Table	Page
1. Predictive Utility of Early Mathematics Screening Studies	6
2. Demographics of Final Participant Sample ($n = 196$).....	16
3. Means and Standard Deviations for Number of Problems Correct for Pilot Data Collection.....	27
4. Means and Standard Deviations for Kindergarten (K) and Grade 1 Measures	30
5. Concurrent Validity: Correlations among Fall Kindergarten Screening and Criterion Measures	31
6. Concurrent Validity: Correlations among Spring Kindergarten Screening and Criterion Measures	32
7. Predictive Validity: Correlations among Fall Kindergarten Screening and Spring Kindergarten Measures.....	33
8. Predictive Validity: Correlations among Fall Kindergarten Screening and Spring Grade 1 Measures.....	34
9. Predictive Validity: Correlations among Spring Kindergarten Screening and Spring Grade 1 Measures.....	35
10. Classification Indices for Logistic Regression Models for MD-Conceptual.....	37
11. Classification Indices for Logistic Regression Models for MD-Operational.....	38

CHAPTER I

INTRODUCTION

Although not explicit in the federal definition (Individuals with Disabilities Education Act Amendments, 1997), an IQ-achievement discrepancy often forms the basis for a learning disability label. This identification procedure is problematic for children in kindergarten or first grade, however, because students in the early grades have not had sufficient exposure to academic curricula to demonstrate such a discrepancy. Further, possible biases in intelligence testing (Valencia & Suzuki, 2001) and the overrepresentation of minority students identified as having a learning disability (Ferri & Connor, 2005) question the validity of this “wait-to-fail” approach (Vaughn & Fuchs, 2003), for younger students as well as older ones. A call for an alternative to the IQ-achievement discrepancy model for identifying learning disability has been issued (e.g., the President’s Commission on Excellence in Special Education, 2001), and a response-to-intervention (RTI) approach represents one possible alternative.

Implementing evidence-based academic interventions and documenting response (or non-response) to these interventions are the major features of RTI (Marston, 2005). Students progress through levels of a prevention system, with increasing intensity, and only those students for whom standard forms of instruction are deemed insufficient receive formal evaluation for placement into special education. Although the Individuals with Disabilities Education Improvement Act (2004) allowed for identification of learning disability within a RTI framework, many questions remain unanswered concerning the standardized, large-scale implementation of this approach (Marston; Mastropieri & Scruggs, 2005).

Regardless of the diagnostic model (i.e., IQ-achievement discrepancy or RTI), accurate assessment of student performance is crucial. Teachers and diagnosticians require reliable and valid measures that document both level of performance and growth. For example, within most RTI models, the main focus of screening (i.e., conducted at one point in time, early in a student's schooling) is to determine which students are at possible risk for academic failure if they do not receive additional intervention. The classroom progress of these students is then monitored with some sort of classroom-based assessment. Trend lines resulting from progress monitoring serve to predict future performance and become the basis for confirming or disconfirming a student's actual risk for academic failure.

Particularly with respect to students in the early grades, measurement tools that screen for the potential risk for developing learning disability represent an important focus of assessment. The earlier risk for future disability is identified, the earlier 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 for young students predicts future reading difficulty (e.g., Berninger, Thalberg, DeBruyn, & Smith, 1987; Kaminski & Good, 1996; National Institute of Child Health and Human Development, 2000; Scarborough, 1998; Torgesen, 1998). Thus, early screening efforts to identify students with such a deficit allow for intervention; the goal is to prevent future reading difficulty. Even so, screening for future reading disability at an early age produces a set of false positives (i.e., students who seem to be at-risk based on the screen, but whose forecasted deficits disappear largely without additional intervention). Nevertheless, the construct of phonemic awareness continues to prove a strong predictor of reading ability.

By contrast, identification of a construct or set of skills that represents a strong predictor of future mathematics difficulty (MD) has yet to be identified. A 2005 issue in the *Journal of Learning Disabilities* focused on the early identification and intervention efforts for students

with (or at risk for) MD. In this issue, Gersten, Jordan, and Flojo (2005) summarized research on early identification for MD. They concluded 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 may offer utility. These skills are both aspects of “number sense” (e.g., Dehaene, 1997; Okamoto & Case, 1996), which may serve as a predictor of mathematics performance for young children.

In contrast to phonemic awareness, which is a language ability that does not involve actual reading, number sense represents actual math knowledge. According to Gersten and Chard (1999), number sense involves the flexibility and ease with which a student mentally computes and intuitively relates mathematical concepts. The authors argued that number sense directly relates to mathematical performance and that screening measures based on this construct should yield predictive information regarding future mathematics ability. As Berch (2005) and Dowker (2005) pointed out, however, *number sense* is not clearly defined or easily operationalized. To illustrate this point, Berch listed 30 alleged components of number sense proposed by various researchers, ranging from “faculty permitting the recognition that something has changed in a small collection when, without direct knowledge, an object has been removed or added to the collection” (No. 1) to “can recognize gross numerical errors” (No. 16) to “process that develops and matures with experience and knowledge” (No. 30). Clearly, number sense means different things to different people. Even so, whether number sense drives arithmetic performance or whether increased arithmetic skill leads to deeper conceptual understanding and stronger number sense remains unknown. 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 an effective means of forecasting which young students are at risk for MD (Gersten et al., 2005). In the meantime, future research should continue to investigate and operationalize the

construct of number sense. Perhaps deficient number sense links directly to MD, with intervention leading to decreased probability of occurrence. Until research more clearly demonstrates the link between specific behaviors indicative of number sense and mathematics outcomes, however, this remains conjecture.

When identifying the type of skills predictive of future mathematics performance, researchers must demonstrate aspects of technical adequacy and predictive utility. With respect to screening measures, adequate reliability of test scores indicates that scores are consistent and reasonably free from measurement error to serve as useful indicators of present level of functioning. Statistics for these indices include a method of rational equivalence such as the Kuder-Richardson formulas or coefficient alpha, the coefficient of stability, and the coefficient of equivalence (Gall, Gall, & Borg, 2003). Additionally, a test's validity is based on the appropriateness of inferences made from the test scores (Salvia & Ysseldyke, 1991). As Cronbach and Meehl (1955) described, test validity can be examined in terms of criterion-related, content, or construct validity. Criterion-related validation can be examined relative to both concurrent and predictive validity (Urbina, 2004) by examining the relationship between the screening measure and valid outcome measures administered at the same time as or a later time frame. A strong correlation suggests the screening measure has tapped the same underlying construct as the criterion measure. With respect to kindergarten mathematics screening instruments, the criterion is future mathematics difficulties. Finally, applying specific criteria to designate risk on the outcome and then comparing the predictions made with actual outcome yields information about the sensitivity and overall accuracy of the screening measure. The predictive utility of a screener represents perhaps the most compelling evidence for the usefulness of a measure in establishing risk status for eventual MD.

Toward that end, researchers investigate the utility of screening young learners for potential MD. In the next section, we summarize prior work assessing MD risk for kindergarten students. We then describe how the present study extends the literature with respect to screening kindergarten students for MD risk and clarify the purposes of this study.

Prior Work Determining MD Risk of Kindergarten Students

We identified 12 studies that targeted kindergarten students, included screening measures or outcome variables specific to mathematics performance and documented the predictive validity or predictive utility of the screening measures (Baker et al., 2002; Bramlett, Rowell, & Mandenberg, 2000; Chard et al., 2005; Clarke, Baker, Smolkowski, & Chard, 2008; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Kurdek & Sinclair, 2001; Lembke & Foegen, 2005; Mazzocco & Thompson, 2005; Pedrotty Bryant, Bryant, Kim, & Gersten, 2006; Simner, 1982; Tiesl, Mazzocco, & Myers, 2001; VanDerHeyden, Witt, Naquin, & Noell, 2001). For each study, Table 1 documents 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, if so provided by the authors (i.e., sensitivity, specificity, and overall accuracy).

Studies that screened children prior to entering kindergarten but did not include evidence of predictive validity or utility or did not include screening measures or outcome variables specific to mathematics performance were excluded. Screening measures for use with children prior to entering kindergarten tend to include more global measures of school “readiness” rather than specific measures of math-related skill (Costenbader, Rohrer, & Difonzo, 2000). Although

Table 1

Predictive Utility of Early Mathematics Screening Studies

<i>Author</i>	<i>n</i>	<i>Grade Screen</i>	<i>Grade Outcome</i>	<i>(Math -Related) Screening Measure(s)</i>	<i>Outcome Measure(s)</i>	<i>Predictive Validity (r)</i>		<i>Predictive Utility</i>		
						(A)	(B)	<i>Sensitivity_y</i>	<i>Specificity_y</i>	<i>Overall Accuracy</i>
Baker et al. (2002)	65, 95	K(S)	1(S)	Number Knowledge Test (NKT)	SAT-9 (A)	.72	.72			
				Digit Span Backward	NKT (B)	.47	.60			
				Numbers from Dictation		.47	.48			
				Magnitude Comparison		.54	.45			
Bramlett et al. (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				
Quantity Discrimination (QD)		.50								
Missing Number (MN)		.64								
Clake et al. (2008)	221	K(F)	K(S)	Oral Counting	SESAT	.55				
				Number Identification		.58				
				Quantity Discrimination		.57				
				Missing Number		.60				
Jordan et al. (2007)	277	K(F)	1(S)	Number Sense Core	WJ Calc + App Prob	.70				
				Counting Skills		.36				
				Number Knowledge		.54				
				Nonverbal Calculation		.52				
				Story Problems		.47				
				Number Combinations		.58				

Table 1 (cont.)

Kurdek & Sinclair (2001)	281	K(F)	4(?)	Kindergarten Diagnostic Inst (KDI)- Form Perception	Ohio Gr 4 State Achievement Test	.27					
						.37					
Lembke & Foegen (2005)	44	K(F)	K(S)	Quantity Discrimination	Teacher Ratings (A)	.64	.33				
						Quantity Array	TEMA-3 (B)	.58	.30		
						Number Identification		.63	.39		
						Missing Number		.44	.41		
Mazzocco & Thompson (2005)	209	K(?)	3(?)	Composite Scores from Various Measures	<10th Percentile on TEMA-		71.4 -	78.2 -	78.7 -		
							91.7%	90.3%	89.4%		
Pedrotty Bryant et al. (2006)	135	K(W)	K(S)	Oral Counting	SAT-10 MPS	.49					
						Number Identification		.51			
						Quantity Discrimination		.61			
						Missing Number		.67			
						Digits Backward		.54			
Simmer (1982)	67	K(F)	K(S)	Writing Reversible Numbers and Letters from STM (form errors)	Gr 1 Readiness (A)						
								-.67	-.40	89.0%	84.2%
Tiesl et al. (2001)	234	K(S)	1(F-S)	Teacher Ratings of Math Level (<10th percentile)	TEMA-2	.34					
								65.2%	87.7%	85.4%	
VanDerHeyden et al. (2001)	25	K(W)	K(S)	Circle Number	Retention						
						Write Number		71.4%	94.4%	88.0%	
						Draw Circles		00.0%	90.9%	80.0%	
					"Validation Problem"		00.0%	91.7%	88.0%		

these measures may answer interesting questions concerning future overall academic performance, they may not specifically predict math performance. Thus, studies that screened pre-kindergarten children with readiness scales (e.g., Augustyniak, Cook-Cottone, & Calabrese, 2004; Kelly & Peverly, 1992) did not meet selection criteria for the purpose of this paper and were read solely for background information. Additionally, although VanDerHeyden et al. (2004) included math-related screening measures with preschool participants, the authors did not examine the predictive utility of the measures. Finally, Magliocca, Rinaldi, and Stephens (1979), for example, studied the efficacy of a screening instrument for identifying at-risk kindergarten and first-grade participants, but did not include predictors or outcome variables specific to math performance. Studies such as these were excluded.

As Table 1 shows, the majority of 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; Tiesl et al., 2001). Only three studies (Jordan et al., 2007; Kurdek & Sinclair, 2001; Mazzocco & Thompson, 2005) allowed for greater than a year to elapse before assessing outcome. (Note: Three studies [Chard et al., 2005; Lembke & Foegen, 2005; Pedrotty Byrant et al., 2006] included samples of both kindergarten and first-grade students; we report results for the kindergarten samples only.)

With the exception of Mazzocco and Thompson (2005) and VanDerHeyden et al. (2001), all studies provided data attesting the predictive validity of their respective screening measures. Correlations ranged from .27 to .72, with an average of .51. Five studies provided information regarding the overall accuracy, sensitivity, and specificity of math screeners, either with predictive validity correlations (Bramlett et al., 2000; Simner, 1982; Tiesl et al., 2001) or without (Mazzocco & Thompson; VanDerHeyden et al.). For these studies, the overall accuracy of the

screeners ranged from 59.8% to 89.4%. Sensitivity ranged widely, from 00.0% to 91.7%; specificity did not range as such (57.5% to 94.4%). Based on these data, screeners were more accurate in 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. Two studies (Bramlett et al., 2000; Simner, 1982) used only one single-skill measure to predict mathematics outcome. Bramlett et al. presented students with randomly ordered numbers (i.e., from 1-20) on a sheet of paper, and students named as many numbers as possible in one minute; Simner had students write the 41 reversible numbers and letters from memory, exposing students to one item at a time for a period of 2.5 seconds. The remainder of the studies with single-skill screening measures used two or more measures to predict math outcome (Baker et al., 2002; Chard et al., 2005; Clarke et al., 2008; Jordan et al., 2007; Kurdek & Sinclair, 2001; Lembke & Foegen, 2005; Pedrotty Bryant et al., 2006; VanDerHeyden et al., 2001); many of the measures used across studies assessed the same skill. For example, the ability to write numbers from dictation was assessed by Baker et al., Chard et al., and VanDerHeyden et al., in addition to Simner. Further, several studies measured students' ability to judge the magnitude of a pair of numbers, i.e., to choose the bigger of two numbers (Baker et al.; Chard et al.; Clarke et al.; Lembke & Foegen; Pedrotty Byrant et al.). Requiring students to state numbers as they were presented visually, identifying the missing number in a sequence of numbers, and counting ability were key skills addressed across several studies, as well.

In contrast to the single-skill screening measures, four studies incorporated multiple-skill screeners to their predictive models. Baker et al. (2002) used the Number Knowledge Test (Okamoto & Case, 1996), an individually administered test 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. The authors selected items from the KeyMath-Revised (KM-R; Connolly, 1998), the Test of Early Mathematics Ability, 2nd Edition (TEMA-2; Ginsburg & Baroody, 1990) the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock & Johnson, 1989) Math Calculations subtest, and the Stanford Binet (4th ed.) (Thorndike, Hagen, & Sattler, 1986) Quantitative Reasoning subtest to assess math abilities. Tiesl et al. (2001) required teachers to rate students' mathematics performance levels with selected items from the Teacher's Report Form (Achenbach, 1991) and the Conners' Teacher Rating Scale (Conners, 1997) short form. Finally, Jordan et al. (2007) combined results from five tasks (i.e., comprising counting skills, number knowledge, nonverbal calculation, story problems and number combinations) to yield a score for "Number Sense Core." Students were assessed across six time points from fall of kindergarten to spring of first grade. Across studies, predictive validity was similar for the single- versus multi-skill screeners. Coefficients for the single-skill screeners ranged from .27 to .67, averaging .54; coefficients for the multi-skill screeners ranged from .36 to .73, with an average of .55. Although some studies used both types of screeners, none specifically tested which type predicted various math outcomes with greater precision, in terms of decision utility.

The majority of studies used outcome variables reflecting mathematics performance on published tests (e.g., the Stanford Achievement Test, 9th ed. [SAT-9; The Psychological Corporation, 1995]; the WJ-R [Woodcock & Johnson, 1989] Calculations and Applied Problems subtests). Yet, authors also reported outcomes such as teacher rankings of kindergarteners' readiness for first grade and June (of first-grade) report card grades in mathematics (Simner, 1982); a teacher rating scale of general math proficiency (Lembke & Foegen, 2005); and professional judgments of academic difficulties (VanDerHeyden et al., 2001). Although some of

these outcomes related to conceptual understanding of mathematics concepts, such as the Number Knowledge Test, or to operational outcomes such as the Calculations subtest of the WJ-R (Woodcock & Johnson), none of the studies specifically addressed whether development could be forecast more precisely for either type of outcome. This seems an important question to address, given the variability in kindergarten classrooms with respect to calculation skill. For example, if kindergarten students are not similarly exposed to curricula that emphasize written computation skills, a screening measure that comprises this skill seems unlikely to generalize across settings. Across studies, predictive validity seemed similar when outcomes such as published tests were used (average of .51) and when outcomes reflected teacher judgment (average of .54). In terms of decision utility data, the sensitivity of screening variables ranged widely, from 0.00% (i.e., VanDerHeyden et al.'s prediction of "Validation Problem") to 91.7% (i.e., Mazzocco & Thompson's 2005 prediction of composite scores on published tests). Authors did not directly address the issue of timed versus untimed mathematics screeners or outcomes in any of the previous studies.

Across these studies, we offer two observations. First, the majority of kindergarten screening studies conducted thus far assessed mathematical outcomes one year or less from the time screening occurred. Because kindergarten students vary in their experience with number concepts prior to commencing formal schooling, assessing math outcome before a substantial amount of mathematics instruction takes place potentially yields an inflated number of false positives. This is problematic in that too many false positives stress the resources available in school settings to provide remediation for students who truly need intervention. Waiting longer than one year before assessing math outcome allows students who have had less preschool exposure to number concepts to "catch up" to their peers via strong classroom instruction, and thus lowers the risk of mistakenly identifying those students as potentially MD.

Our second observation concerns the predictive utility of kindergarten math screening tests. The majority of studies we reviewed relied on predictive validity correlational data as an indication of a measure's ability to predict students' risk for developing MD. Few studies, however, looked beyond predictive correlations to evaluate the sensitivity or specificity of math screeners. Although predictive correlations do provide a certain amount of support for the value of a kindergarten screening event, the decision utility data that could further attest a screener's value are missing from the majority of previous work.

How the Present Study Extends Previous Work

In the present study, we sought to extend previous work on early math screening in several ways. First, by piloting the screening tests, we allowed for item response theory analyses to order the items by difficulty, eliminate items with poor discrimination, and establish an administration ceiling for the untimed portion of the assessment. This increased efficiency of administration. Second, we adopted a longer perspective than in most prior studies, screening the students in the fall and spring of kindergarten and subsequently retesting during the spring of first grade to investigate the accuracy of the screening measures in identifying students who develop math difficulties in first grade. By contrast, the majority of studies we reviewed allowed for one year or less of elapsed time before assessing student outcome. Third, in addition to providing evidence of the technical adequacy (i.e., reliability; concurrent and predictive validity) of the screeners, we also examined the math screeners' predictive utility with respect to sensitivity and specificity. Few of the studies we reviewed provided this information. Finally, and in a related way, we extended previous research on the predictive utility of kindergarten math screeners by evaluating (a) the predictive accuracy of single- versus multi-skill screeners, (b) fall versus spring administration of kindergarten testing, and (c) conceptual versus operational outcomes. To our knowledge, no previous studies have addressed these specific

questions that shed light on the benefit of single- versus multiple-skill screening measures, the most opportune time in the kindergarten year to screen for MD (i.e., fall vs. spring), and whether conceptual or operational mathematics skill should be the focus of outcome.

Our research questions included the following: What is the reliability of mathematics screening measures for kindergarten students? What are the concurrent and predictive validities of these measures, with respect to kindergarten and grade one performance on the EMDA (The Psychological Corporation, 2002a), the Estimation and Numerations subtests of the KM-R (Connolly, 1998), and First-Grade Math CBM Computation and Concepts/Applications (Fuchs, Hamlett, & Fuchs, 1989; 1990)? How do single-skill versus multiple-skill math screeners compare in terms of predictive efficiency? How accurate is fall versus spring kindergarten screening? And finally, Can first-grade mathematics development be forecast more precisely in terms of conceptual or operational outcomes?

CHAPTER II

METHOD

Participants

Twenty kindergarten teachers from five schools in a southeastern metropolitan school district were randomly selected from a pool of interested teachers to participate in the study. Two schools each had three participating teachers, two additional schools each had four participating teachers, and the remaining six kindergarten teachers were from one school. Ten of the 20 kindergarten classrooms received Title-1 funding due to low socio-economic status of the enrolled student population in the school. From the 20 classrooms, 252 students returned signed parental consent and participated in the initial testing wave in the fall of kindergarten. Of the original 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 two years of the study (i.e., 20 students moved out of the school district before the end of the first year of the study, and 36 additional students moved during the second year). We used inferential statistics to compare the students who exited versus those who remained on demographic variables and screening scores. There were no significant differences except on the Number Sense multi-skill screener. The mean score of students who exited the study on this measure was 12.91 ($SD = 6.04$); for those who remained, 15.65 ($SD = 6.80$). In this study, we report results for the sample of 196 students with complete data.

Participating teachers provided demographic information on consented students' date of birth, gender, subsidized lunch status, race, special education status, English language learner status, previous preschool experience, math ranking, and minutes of daily math instruction (i.e., by classroom). The average age of students at the onset of the study was 5 years 8 months (however, two students did not provide this data). Students received, on average, 49.08 minutes of daily math instruction ($SD = 20.83$). See Table 2 for the remaining demographic information for the sample.

Kindergarten Screening Measures

Two of the kindergarten math tests were multiple-skill screeners: *Computation Fluency*, which is group administered, and *Number Sense*, which is individually administered. Items for inclusion were determined from three sources: (a) from interviews with experienced kindergarten and first-grade teachers; (b) from examination of the existing literature base and the published set of kindergarten academic standards of the school district in which the study took place; and (c) from discussions with university professors familiar with elementary school kindergarten skills. Further, 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 to order the items by difficulty and devise a ceiling rule for the administration of the individually administered screening measure. The ceiling rule allowed examiners to discontinue testing after five consecutive incorrect answers, shortening the assessment time for some students.

Computation fluency. The first multi-skill measure, Computation Fluency, is a 5-min timed assessment of counting, addition, and subtraction fluency. It is administered in a

Table 2

Demographics of Final Participant Sample (n = 196)

<i>Variable</i>	<i>n</i>	<i>%</i>
Males	103	52.55
Subsidized Lunch	101	51.53
Race: African American	71	36.22
Caucasian	86	43.88
Hispanic	21	10.71
Asian	11	5.61
Kurdish	4	2.04
Other ^a	3	1.53
Special Ed Diagnosis: None	170	86.73
Learning Disability	1	0.51
Speech/Language	12	6.12
Gifted	11	5.61
Other ^b	2	1.02
English Language Learner	9	4.59
Known to Attend Preschool	99	50.50
Teacher Math Rating: Above Grade Level	49	25.00
Grade Level	116	59.20
Below Grade Level	31	15.80

Note: ^aOne student each was Indian, Samolian, or Iraqi. ^bOne student each was diagnosed as having a Visual Impairment or Developmental Delay.

whole-class setting and includes 25 items (five items each of five problem types) presented randomly 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). This measure contains five rows of five problems each; the items are bordered in black to help delineate each problem. The examiner conducts a scripted 10-min whole-class lesson explaining how students respond to the five types of items and that they need to stop working (i.e., pencil held in the air) when the timer goes off. After this brief administration lesson, the examiner instructs students to answer as many problems as they can, to look for the easiest problems first, and then to go back to try the harder ones. The student is not penalized for number reversals or poorly formed written responses. Scores of correct responses (across the five types of items) in 5 min are recorded. We created two forms, identical in format but comprising different items.

Computation Fluency is conceptually based on the Computation CBM probes for grades one through six as developed by Fuchs and colleagues (e.g., Fuchs, Fuchs, Hamlett, Phillips, et al., 1994; Fuchs & Fuchs, 2004). It resembles the Computation CBM probes in appearance; both Fuchs's CBM probes and the Computation Fluency subtest include five rows of five items in a bordered grid design. Further, it samples computation items across the kindergarten curriculum, as do the CBM probes for grades one through six sample computation items for the corresponding grade level curriculum. Because it can be group administered, is brief in duration (i.e., 5 min), and easily scored, this measure has potential for use as screening and progress monitoring, as are the CBM probes at the higher grade levels. See Appendix A for an example of the Computation Fluency measure.

Number sense. The second multi-skill measure, Number Sense, is individually administered. It samples a greater number of mathematics skills at the kindergarten level, with 30 items (3 items each of 10 types), ordered in difficulty from easiest to hardest, based on item response analyses of the pilot data. The 10 types of items are quantity discrimination, mental number lines, ordering numbers, estimation, patterns, counting backward, shape discrimination, number sentences, writing numbers, and one-to-one correspondence. The tester reads the directions from a script for each item to the student, and then allows up to 1 min for the student to respond or moves on as soon as the child responds. The student is provided a pencil and writes answers to items; as with Computation Fluency, the student is not penalized for misspelled or poorly formed written responses. The five pages of this measure each contain six items; the examiner holds a piece of cardstock over the items and slides the cardstock down to expose one new item at a time. The examiner scores each item immediately following the student's response. Correct responses receive a score of 1; incorrect responses receive a score of 0. The examiner stops administering items after five consecutive scores of 0. The score is the number of correctly answered items.

Number Sense, similar to the Concepts/Applications CBM probes developed by Fuchs and colleagues (Fuchs & Fuchs, 2004; Fuchs, Hamlett, & Fuchs, 1989), is a multiple-skill screener that samples grade-level skills. However, it differs from the Concepts/Applications CBM probes in that it is not designed for group administration, items are scored immediately subsequent to each response, and a ceiling rule limits the length of the test for some students. See Appendix B for a copy of the Number Sense measure and Appendix C for the scoring sheet.

In the spring of 2005, Computation Fluency and Number Sense were piloted with 90 kindergarten students in three public elementary schools. All three schools received Title-1 funding; 46 (i.e., 51.1%) of the students in the pilot sample were female; 53 (i.e., 58.9%) of the

students were six years old at the time of testing (all others were five years old). Interscorer agreement was computed with 18 (i.e., 20%) of both the Computation Fluency and Number Sense protocols. A second scorer independently scored the 36 total tests; interscorer agreement for each subtest was calculated as the number of tests for which both scorers agreed on the score divided by 18. Interscorer reliability was .94 for Computation Fluency and 1.00 for Number Sense.

Students' average score on the Computation Fluency subtest was 13.77 ($SD = 5.78$) of a possible score of 25. Further, the data from the pilot group showed a normal distribution of scores that corresponded with ability level, indicating that individual differences in computation skill could be indexed with this measure. The average score for the pilot group on the Number Sense subtest was 18.98 ($SD = 5.96$) of a possible score of 30. These data similarly demonstrated a normal distribution of scores and slight negative skewness. See Table 3 for means and standard deviations for all classrooms in the pilot study. All teachers provided their students' scores on a district-mandated kindergarten test (administered by the teacher during the same time frame) to allow comparison with the screening measures. The district test correlated .64 with Computation Fluency and .75 with Number Sense; the Computation Fluency measure correlated .69 with Number Sense. Coefficient alpha for this pilot study sample was .88 for Computation Fluency and .87 for Number Sense.

Quantity discrimination. The third and single-skill kindergarten screening measure, Quantity Discrimination (QD; Chard et al., 2005), is a 1-min timed probe measuring students' ability to name the larger of two numbers (ranging from 0 to 10), presented in 28 individual boxes across two pages. Clarke et al. reported test-retest reliability as .85-.99 and concurrent and predictive validity coefficients that ranged from .70 to .80. The QD measure was chosen because

it has demonstrated strong predictive capability for early mathematics skill (Clarke & Shinn, 2004) for first graders and strong predictive capabilities for kindergartners (Chard et al).

Outcome Measures and MD Designation

Early math diagnostic assessment math reasoning and numerical operations. The EMDA (The Psychological Corporation, 2002a) is an individually-administered norm-referenced test for use with preschool to third-grade students. The test, which takes approximately 20 min to administer, 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. Students are shown a stimulus page corresponding to each item and orally respond to the examiner's prompts. Numerical Operations measures one-to-one correspondence, number identification, number writing, calculation, and rational numbers. Students identify and circle numbers within a mixed set of numbers and letters; write numbers as prompted by the examiner; count a set of eight pennies and write the amount; and write answers to arithmetic computation problems. The items are ordered by difficulty, and basal and ceiling rules are provided. The test yields raw scores, percentile ranges, and standard scores. The EMDA examiner's manual provides reliability coefficients ranging from .71 to .93. Correlations with the Wechsler-Individual Achievement Test (The Psychological Corporation, 1992b) are listed in the manual as .82 and .78, correlations with the Wide Range Achievement Test-Revised (Wilkinson, 1993) as .67 and .77. The EMDA was selected for its appropriateness with young children, its ease of administration (i.e., advance degree not required), and its inclusion of skills similar to those of the screening measure.

Keymath-revised numeration and estimation. The KM-R (Connolly, 1998) is an individually administered norm-referenced test for use with students from kindergarten through grade 12. Two subtests were used in this study: Numeration (i.e., concepts such as counting,

correspondence, sequencing numbers, and ordinal positions) and Estimation (i.e., estimation of rational numbers, measurement, and computation). As with the EMDA, test items are ordered by difficulty, basal and ceiling rules are provided, and raw scores, standard scores, and percentile ranks are available. The examiner's manual reports alternate form reliability coefficients as .50 to .70 for the subtests and .90 for the entire test. 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 reported as .67 and .43, respectively. The KM-R was selected for similar reasons as the EMDA; in addition, it was selected because it provides a measure of estimation.

CBM computation and concepts/applications. At the end of first grade (i.e., the second year of the study), we assessed participating students with First-Grade Computation and Concepts/Applications CBM probes (Fuchs & Fuchs, 2004; Fuchs, Hamlett, & Fuchs, 1989), which sample items from the first-grade curriculum. These items are presented to students in a 25 item 3-min timed test for Computation and in a 22 item (approximately) 10-min test for Concepts and Applications. Each CBM test is scored as number of problems and number of digits correct. Each alternate form of each test contains a comparable number of items representing the same group of problem types, and data from these probes provide the basis of progress monitoring over time.

MD designation. Students received a designation of MD in one of two ways: scoring below the 16th percentile on either the EMDA Math Reasoning subtest or the EMDA Numerical Operations subtest at the end of first grade (or the end of the second year of kindergarten, if a student repeated kindergarten). We used the normative tables provided by the examiner's manual for designating MD.

Interscorer Agreement

Data were examined for interscorer agreement at each of three testing waves. After the first wave of testing (i.e., fall of kindergarten), a second scorer independently scored approximately 20% of all protocols. Interscorer agreement (computed by dividing the number of agreed points by the total number of points, across tests) ranged from 99.29 – 100.0%. This procedure was repeated after the second testing wave (i.e., spring of kindergarten). Interscorer agreement at this wave ranged from 98.96 – 100.0%. Following the third testing wave (i.e., spring of first grade), 100% of the testing protocols were rescored by a second scorer for accuracy, and all discrepancies were resolved by examining the original products.

Procedure

Participating students were tested by the first author and by trained examiners. All examiners were graduate students with varying degrees of classroom experience; trained to acceptable levels of accuracy during practice sessions; and monitored by the first author throughout all testing waves.

We administered tests to students in three waves. During the first wave (i.e., fall of kindergarten), students were tested on three separate days. On the first day, students received one form of Computation Fluency in a whole-class setting as well as the individually administered Number Sense subtest. One-half of the students were randomly chosen to receive Form A of Computation Fluency; the other half, Form B. One week later, students were tested with both subtests of the EMDA and both subtests of the KM-R. The following week (i.e., two weeks had elapsed from the first day of testing), students received the alternate form of Computation Fluency; however, this time, it was administered on an individual basis. Students also received QD following the administration of Computation Fluency.

During the second testing wave (i.e., the final weeks of kindergarten), students were again tested across three weeks and on three separate days. The testing schedule was identical to

that of the first wave, with one exception: Both administrations of Computation Fluency were group-administered.

The third testing wave occurred during the final weeks of the subsequent school year. For most students, this was the end of first grade. However, three students repeated kindergarten, so this wave occurred at the end of their second full year of kindergarten. At this point, students had dispersed from 20 classrooms in five public schools to 45 classrooms in 22 public schools and two local-area private schools. In the fall of this school year, parents received a letter reminding them of their consent and apprising them that their child(ren) would be tested again in the spring, for follow-up purposes. Teachers of these students were also contacted to schedule convenient testing times.

As with the previous two testing waves, assessment occurred over three weeks and on three separate days. On the first day, students received one form of CBM Computation and CBM Concepts/Applications tests. One week later, testers administered the EMDA subtests and the KM-R Numeration subtest. (Because of a floor effect for the KM-R Estimation subtest when administered the previous times, and because one of the examiners administered this subtest incorrectly to a large group of students in the previous testing wave, we elected to omit this test from the final testing wave.) Finally, testers returned the following week to administer alternate forms of the first-grade CBM tests. All testing was conducted individually at this wave.

Data entry was conducted by two graduate students independently into two separate, but identical, Excel spreadsheets. The databases were compared for discrepancies, which were resolved by examining the original protocols. In this way, a final spreadsheet was created and imported into SPSS 16 for analyses.

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 multi-skill screeners and alternate form reliability (i.e., Pearson product moment correlation coefficients for Forms A and B) of Computation Fluency.

Correlations among screening and outcome measures. We examined the concurrent validity of the three kindergarten screening measures (i.e., Quantity Discrimination, Computation Fluency, and Number Sense) 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, KM-R subtest, and CBM mathematics tests.

Logistic regression to predict MD. We used logistic regression to evaluate the utility of the kindergarten screening measures for predicting MD status, separately for math reasoning (i.e., conceptual) and numerical operations (i.e., operational) outcomes. Binary logistic regression is used when the outcome variable is dichotomous (e.g., MD vs. not-MD); predictor variables (e.g., scores on the screeners) can be of any type. Logistic regression provides the percentage of variance in the outcome variable that is explained by the predictor variable(s), as well as a ranking of the independent variables' relative importance. The output of a logistic regression analysis is a set of equation coefficients that allows for the calculation of the probability that a case is of certain class. Logistic regression is used rather than linear regression when the outcome is binary because logistic regression does not assume a linear relationship

between the predictor and outcome variables; normal distribution of the outcome variables or error terms; homogeneity of variance; or interval-level or unbounded predictor variables.

Within the context of RTI, we were most interested in maximizing the number of students who truly required additional and intensive mathematics instruction (i.e., “true positives”) while limiting the number of those who did not (i.e., “false positives”). The set of true and false positives would comprise the set of students identified for secondary intervention. For this reason, we set the classification cutoff for the logistic regression models to be equal to the proportion of first-grade MD children in the sample. We used SPSS 16.0 statistical software to generate the logistic regression models, and entered the screeners independently to contrast their predictive capabilities.

ROC curves to contrast various models. We used measures of sensitivity, specificity, overall hit rate, and area under the ROC curve (AUC) to contrast the utility of various logistic regression models. First, sensitivity refers to the true positives, that is, the proportion of children correctly predicted by the model to be MD (in this study). Sensitivity is computed by dividing the number of true positives by the sum of true positives and false negatives. Second, specificity, or true negatives, by contrast, represents the proportion of children correctly predicted to be *not* MD. Specificity is computed by dividing the number of true negatives by the sum of true negatives and false positives. Third, the overall hit rate refers to the proportion of children correctly classified as either MD or not-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 cutpoints 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 this procedure, imagine that we had already placed children into their correct MD or not-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 not-MD group. The AUC represents the proportion of randomly chosen pairs of students for which the screeners correctly classified as MD versus not-MD. It ranges from .50 to 1.00. The greater the AUC, the less likely that classification was 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, Fuchs, Compton, Bryant, Hamlett, & Seethaler, 2007). The output from ROC analyses includes confidence intervals for the AUC and a lack of overlap for the confidence intervals across models indicates significant difference in predictive accuracy for the models.

CHAPTER III

RESULTS

Descriptive Statistics

See Table 3 for the means and standard deviations of each test for each of the three testing waves.

Table 3

Means and Standard Deviations for Number of Problems Correct for Pilot Data Collection

	K-Math Test				
	<i>n</i>	Computation Fluency ^a		Number Sense ^b	
		M	(SD)	M	(SD)
Class #1	18	11.83	(5.02)	18.00	(5.35)
Class #2	11	10.45	(4.41)	15.36	(6.48)
Class #3	13	14.23	(6.62)	20.23	(5.60)
Class #4	15	16.60	(6.01)	19.53	(6.36)
Class #5	16	15.63	(5.32)	22.25	(4.97)
Class #6	17	13.64	(5.23)	18.36	(5.80)
Overall	90	13.77	(5.78)	18.98	(5.96)

Note: ^a number correct out of 25 items; ^b number correct out of 30 items.

One purpose of this study was to evaluate the technical adequacy of the kindergarten screening measures. With respect to reliability of the scores, we evaluated inter-item consistency of both Computation Fluency and Number Sense with coefficient alpha, and content sampling consistency of the alternate forms of Computation Fluency. Because previous work had evaluated the reliability of the single-skill, Quantity Discrimination measure (e.g., Chard et al., 2005; Clarke & Shinn, 2004; Lembke & Foegen, 2006; Pedrotty Bryant et al., 2006), we were interested in the reliability of only the two multi-skill screeners.

We evaluated inter-item consistency for the fall administration of Computation Fluency as follows. Students received two forms of the measure (i.e., Forms A and B). Half of the students were randomly selected to receive Form A during the first (group) administration and Form B during the second (individual) administration; the remaining students received first Form B 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, Computation Fluency was administered in a group format at both occasions. In this way, alpha for the fall administration of Computation Fluency averaged .88 and for the spring administration averaged .92. For the same set of students, coefficient alpha for Number Sense was .91 for the fall administration and .88 for the spring.

Alternate form reliability for Computation Fluency was determined by correlating each student's score on Form A with his or her score on Form B. In the fall and spring testing occasions of kindergarten, correlations were significant and .54 and .77, respectively. Note that tests were administered both within a group and individually in the fall; by contrast, in the spring, all tests were group administered. To evaluate the degree to which the fall group and fall individual testing administration formats were related, we also examined the correlation between

students' scores as a function of testing format; the scores correlated at a statistically significant .72.

To further examine the technical adequacy of the kindergarten math screeners, we examined the concurrent and predictive validity of the scores with various mathematics outcome measures. With respect to concurrent validity, Table 4 provides the zero-order correlations for the fall kindergarten screening and criterion measures; Table 5 provides the same information for the second wave of testing (i.e., spring of kindergarten). All correlations at both testing occasions were significant at the 0.01 (2-tailed) level. With the exception of correlations with the KM-R Estimation subtest, which ranged from .26 to .32 in the fall and from .35 to .41 in the spring, correlations for the kindergarten screeners with outcome measures ranged from .60 to .79 in the fall and from .55 to .74 in the spring.

Similar to the concurrent validity correlations, all predictive validity correlations were significant at the 0.01 (2-tailed) level. See Tables 6, 7, and 8 for 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. For the first set of test data (i.e., fall of kindergarten with spring of kindergarten measures), correlations ranged from .53 to .82, excluding those with KM-R Estimation, which ranged from .34 to .49. Furthermore, the predictive validity data were similar for all three kindergarten screeners with the math outcome measures. Regarding the predictive validity for the spring of first-grade math outcomes, there was not much difference in range for the fall versus spring kindergarten testing occasions. As Tables 7 and 8 show, predictive validity correlations ranged from .43 to .72 when

Table 4

Means and Standard Deviations for Kindergarten (K) and Grade 1 Measures

<i>Measures</i>	Grade K Fall				Grade K Spring				Grade 1 Spring			
	<i>M^a</i>	<i>(SD^a)</i>	<i>M^b</i>	<i>(SD^b)</i>	<i>M^a</i>	<i>(SD^a)</i>	<i>M^b</i>	<i>(SD^b)</i>	<i>M^a</i>	<i>(SD^a)</i>	<i>M^b</i>	<i>(SD^b)</i>
CF1 (K Fall: Group)	7.55	(5.12)	-	-	16.27	(6.12)	-	-	-	-	-	-
CF 2 (K Fall: Ind)	11.22	(5.72)	-	-	17.58	(6.14)	-	-	-	-	-	-
CF Avg	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)	-	-
CBM C/A, Form 2	-	-	-	-	-	-	-	-	20.31	(4.70)	-	-
CBM C/A, Average	-	-	-	-	-	-	-	-	20.77	(4.16)	-	-

Note: n = 196. ^a Raw score. ^b Standard score. CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; CF Avg = average score of CF 1 and CF 2; NS = Number Sense; KM-R Num = KeyMath-Revised Numeration subtest; KM-R Est = KM-R Estimation subtest; EMDA MR = Early Mathematics Diagnostic Assessment Math Reasoning subtest; EMDA NO = EMDA Numerical Operations subtest; QD = Quantity Discrimination; CBM Comp = Grade 1 Curriculum-based Measurement Computation probe; CBM C/A = Grade 1 CBM Concepts and Applications probe.

Table 5

Concurrent Validity: Correlations among Fall Kindergarten Screening and Criterion Measures

	CF1	CF2	CFAvg	NS	QD	KM-R Num	KM-R Est	EMDA MR	EMDA NO
CF1	--								
CF2	.72	--							
CFAvg	.92	.94	--						
NS	.58	.67	.68	--					
QD	.55	.67	.66	.71	--				
KM-R Num	.55	.59	.62	.67	.64	--			
KM-R Est	.26	.29	.30	.30	.31	.32	--		
EMDA MR	.60	.68	.69	.79	.66	.67	.39	--	
EMDA NO	.56	.59	.62	.68	.60	.61	.26	.62	--

Note: All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; CFAvg = averaged score of CF1 and CF2; NS = Number Sense; QD = Quantity Discrimination; KM-R Num = KeyMath-Revised, Numeration subtest; KM-R Est = KM-R Estimation subtest; EMDA MR = Early Math Diagnostic Assessment, Math Reasoning subtest; EMDA NO = EMDA Numerical Operations subtest.

Table 6

Concurrent Validity: Correlations among Spring Kindergarten Screening and Criterion Measures

	CF 1	CF 2	CFAvg	NS	QD	KM-R Num	KM-R Est	EMDA MR	EMDA NO
CF 1	--								
CF 2	.79	--							
CFAvg	.94	.95	--						
NS	.67	.69	.72	--					
QD	.61	.64	.66	.68	--				
KM-R Num	.62	.60	.64	.68	.61	--			
KM-R Est	.35	.34	.36	.38	.34	.41	--		
EMDA MR	.71	.68	.74	.74	.64	.68	.49	--	
EMDA NO	.64	.62	.67	.55	.56	.58	.40	.66	--

Note: All correlations significant at the 0.01 level (2-tailed). CF1 = Computation Fluency, first administration; CF2 = Computation Fluency, second administration; CFAvg = averaged score of CF1 and CF2; NS = Number Sense; QD = Quantity Discrimination; KM-R Num = KeyMath-Revised, Numeration subtest; KM-R Est = KM-R Estimation subtest; EMDA MR = Early Math Diagnostic Assessment, Math Reasoning subtest; EMDA NO = EMDA Numerical Operations subtest.

Table 7

Predictive Validity: Correlations among Fall Kindergarten Screening and Spring Kindergarten Measures

<i>Fall Kindergarten</i>	<i>Spring Kindergarten</i>								
	CF1	CF2	CFAvg	NS	KM-R Num	KM-R Est	EMDA MR	EMDA NO	QD
CF1	.58	.52	.58	.54	.58	.48	.61	.51	.49
CF2	.67	.62	.67	.64	.66	.44	.68	.57	.62
CFAvg	.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; CFAvg = average score of CF1 and CF2; NS = Number Sense; KM-R Num = KeyMath-Revised, Numeration subtest; KM-R Est = KeyMath-Revised, Estimation subtest; EMDA MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest; EMDA NO = EMDA Numerical Operations subtest; QD = Quantity Discrimination.

Table 8

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

Fall Kindergarten	<i>Spring Grade 1</i>								
	KM-R Num	EMDA MR	EMDA NO	CBM1	CBM2	CBMAvg	C/A1	C/A2	C/AAvg
CF1	.58	.59	.56	.41	.45	.46	.42	.44	.46
CF2	.64	.65	.53	.45	.48	.50	.50	.50	.54
CFAvg	.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; CFAvg = average score of CF1 and CF2; NS = Number Sense; Num = KeyMath-Revised, Numeration subtest; MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest; NO = EMDA Numerical Operations subtest; CBM1 = Gr 1 Curriculum-based measurement Computation probe, first administration; CBM2 = second administration; CBMAvg = average score of CBM1 and CBM2; C/A1 = Gr 1 Concepts and Applications probe, first administration; C/A2 = second administration; C/Aavg = average score of C/A1 and C/A2; QD = Quantity Discrimination.

Table 9

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

<i>Spring Kindergarten</i>	<i>Spring Grade 1</i>								
	KM-R Num	EMDA MR	EMDA NO	CBM1	CBM2	CBMAvg	C/A1	C/A2	C/AAvg
CF1	.60	.66	.59	.51	.53	.56	.55	.58	.61
CF2	.59	.62	.51	.45	.51	.52	.53	.55	.58
CFAvg	.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; CFAvg = average score of CF1 and CF2; NS = Number Sense; Num = KeyMath-Revised, Numeration subtest; MR = Early Mathematics Diagnostic Assessment, Math Reasoning subtest; NO = EMDA Numerical Operations subtest; CBM1 = Gr 1 Curriculum-based measurement Computation probe, first administration; CBM2 = second administration; CBMAvg = average score of CBM1 and CBM2; C/A1 = Gr 1 Concepts and Applications probe, first administration; C/A2 = second administration; C/AAvg = average score of C/A1 and C/A2; QD = Quantity Discrimination.

using the fall kindergarten test scores; from .44 to .72 when using the spring kindergarten test scores (i.e., using the averaged scores of the two forms of Computation Fluency).

MD Prevalence as a Function of Mathematics Outcome

We determined MD prevalence for students based on their performance on criterion measures administered at the third testing wave, that is, the end of first grade. This allowed for approximately two 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 either the EMDA Math Reasoning subtest or the EMDA Numerical Operations subtest. The former focused primarily on conceptual skills and mental manipulation of whole numbers; students scoring below the 16th percentile on this subtest were designated MD-conceptual. In contrast, the EMDA Numerical Operations subtest measured students' ability to identify numerical symbols and perform written calculations; students scoring below the 16th percentile on this subtest were designated MD-operational. Based on these criteria, 40 students (i.e., 20.41% of the sample) were MD-conceptual and 59 students (i.e., 30.10%) were MD-operational. Twenty-one students (i.e., 10.71%) met criteria for both MD designations.

ROC Curves to Contrast the Predictive Utility of Logistic Regression Models

In Tables 10 and 11, we report the results of the logistic regression analyses for predicting MD status at the end of first grade, with respect to conceptual and operational outcomes. The tables show the predictive utility of the three kindergarten math screeners when administered to students in the fall and in the spring. Hit rate (i.e., overall accuracy), sensitivity, specificity, and area under the ROC curve (AUC) are included for each math screener.

For predicting MD-conceptual based on the fall-administered screeners (i.e., the top half of Table 10), the single-skill Quantity Discrimination measure resulted in a hit rate of 74.5%,

Table 10

Classification Indices for Logistic Regression Models for MD-Conceptual

<i>Outcome/Model</i>	<i>B</i>	<i>SE</i>	<i>Wald</i>	<i>p</i>	<i>TN</i>	<i>FN</i>	<i>TP</i>	<i>FP</i>	<i>Hit Rate</i>	<i>Sens</i>	<i>Spec</i>	<i>ROC</i>		
												<i>AUC</i>	<i>SE</i>	<i>CI</i>
<u>Fall Predictors</u>														
Quantity Discrimination	-.206	.037	30.233	.000	113	7	33	43	74.5	82.5	72.4	.857	0.03	.797-.916
Constant	1.042	.386	7.288	.007										
Computation Fluency (ind)	-.245	.049	24.919	.000	108	7	33	48	71.9	82.5	69.2	.797	.033	.732-.862
Constant	.912	.432	4.448	.035										
Number Sense	-.207	.035	35.007	.000	121	8	32	35	78.1	80.0	77.6	.841	.030	.783-.900
Constant	1.377	.446	9.525	.002										
<u>Spring Predictors</u>														
Quantity Discrimination	-.168	.026	40.187	.000	126	9	31	30	80.1	77.5	80.8	.861	.035	.793-.929
Constant	2.303	.548	17.649	.000										
Computation Fluency	-.276	.045	37.657	.000	116	8	32	40	75.5	80.0	74.4	.860	.028	.806-.915
Constant	2.655	.611	18.890	.000										
Number Sense	-.315	.051	37.416	.000	124	7	33	32	80.1	82.5	79.5	.877	.028	.822-.931
Constant	4.887	.986	24.544	.000										

Table 11

Classification Indices for Logistic Regression Models for MD-Operational

<i>Outcome/Model</i>	<i>B</i>	<i>SE</i>	<i>Wald</i>	<i>p</i>	<i>TN</i>	<i>FN</i>	<i>TP</i>	<i>FP</i>	<i>Hit Rate</i>	<i>Sens</i>	<i>Spec</i>	<i>ROC</i>		
												<i>AUC</i>	<i>SE</i>	<i>CI</i>
<u>Fall Predictors</u>														
Quantity Discrimination	-.074	.018	16.314	.000	82	21	38	55	61.2	64.4	59.9	.690	.040	.612-.768
Constant	.268	.298	.808	.369										
Computation Fluency (ind)	-.102	.031	10.627	.001	76	25	34	61	56.1	57.6	55.5	.639	.041	.558-.720
Constant	.237	.350	.456	.499										
Number Sense	-.110	.025	19.028	.000	96	25	34	41	66.3	57.6	70.1	.696	.040	.619-.774
Constant	.775	.388	3.987	.046										
<u>Spring Predictors</u>														
Quantity Discrimination	-.062	.017	14.105	.000	95	26	33	42	65.3	55.9	69.3	.661	.043	.577-.745
Constant	.701	.426	2.705	.100										
Computation Fluency	-.136	.030	21.114	.000	89	24	35	48	63.3	59.3	65.0	.722	.037	.649-.794
Constant	1.343	.484	7.703	.006										
Number Sense	-.130	.030	18.164	.000	87	24	35	50	62.2	59.3	63.5	.687	.041	.605-.768
Constant	1.914	.655	8.551	.003										

with sensitivity (82.5%) exceeding specificity (72.4%). The multi-skill screeners, Computation Fluency and Number Sense, resulted in similar fashion. Hit rates for those screeners were 71.9% and 78.1%, respectively, and sensitivity for both (82.5% and 80.0%) exceeded specificity (69.2% and 77.6%). The AUCs for the three fall screeners were .857, .797, 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, 7 to 8 students who were designated MD-conceptual were missed (i.e., see “FN” column) and 35 to 48 students who were identified with the screeners as at risk did not meet end-of-first-grade criterion for MD-conceptual (i.e., see “FP” column).

For predicting the same MD-conceptual outcome, yet based on the spring-administered screening measures (i.e., the bottom half of Table 10), similar results were found. The single-skill and multi-skill screeners resulted in hit rates ranging from 75.5% (Computation Fluency) to 80.1% (both Quantity Discrimination and Number Sense). Quantity Discrimination resulted in higher specificity (80.8%) than sensitivity (77.5%); the multi-skill Computation Fluency and Number Sense showed the reverse, with sensitivity (80.0% and 82.5%, respectively) exceeding specificity (74.4% and 79.5%, respectively). AUCs ranged from .860 to .877, which are deemed good, and overlapping confidence intervals again attested to statistical equivalence across models. False negatives ranged from 7 to 9 with the spring administration of the screeners; false positives ranged from 30 to 40.

For predicting MD-operational status, the three screeners performed similarly in the fall and in the spring (see Table 11). Hit rates for Quantity Discrimination, Computation Fluency, and Number Sense based on fall screening were 61.2%, 56.1%, and 66.3%, respectively. Based on spring screening, the hit rates changed only slightly: 65.3%, 63.3%, and 62.2%, respectively. Sensitivity across both testing occasions ranged from 57.6% to 64.4%; specificity ranged from

55.5% to 70.1%. With the exception of the spring-administered Computation Fluency, which resulted in an AUC of .722 (deemed fair), the screeners' AUCs were all less than .70 (deemed poor). Number of false negatives (i.e., missed students) ranged from 21 to 26 and number of false positives ranged from 41 to 61. The predictive utility of the three screening measures were statistically equivalent at both kindergarten testing occasions, based on overlapping confidence intervals of their corresponding AUCs.

Although there were no significant differences when looking separately at MD-conceptual and MD-operational results (i.e., screeners performed similarly, irrespective of testing occasion, when predicting MD-conceptual or MD-operational status), there was a significant difference when combining the results. Specifically, the screeners predicted future MD status in terms of conceptual outcome with significantly greater accuracy than in terms of operational outcome. The AUCs for the three screeners when predicting MD-conceptual were higher than when predicting MD-operational; their non-overlapping confidence intervals indicated statistical significance.

CHAPTER 1V

DISCUSSION

We evaluated the technical adequacy and predictive utility of one single-skill and two multi-skill measures for screening kindergarten students for risk for MD. The single-skill screener assessed students' ability to discriminate larger numbers from pairs of numbers ranging from 0-10 in one minute. The multi-skill screeners assessed computational fluency and various mathematical concepts central to typical early mathematical development. Conceptual and operational math outcomes were assessed at the end of first grade, with MD operationalized as performance below the 16th percentile on nationally norm-referenced tests.

Previous studies had investigated the reliability and validity of the single-skill (i.e., Quantity Discrimination) screening measure (Chard et al., 2005; Clarke & Shinn, 2004; Lembke & Foegen, 2006; Pedrotty Bryant et al., 2006). Results from these earlier studies showed reliability, on average, to be about .90, with concurrent and predictive validity averaging approximately .60. Our results echo these findings with respect to validity. We found average validity correlations for this test to range from .57 to .63 with criterion measures (i.e., excluding the KM-R Estimation scores, for reasons mentioned previously). With the present study, we focused our attention on the technical adequacy of the two multi-skill kindergarten screeners (i.e., Computation Fluency and Number Sense), even as we considered the validity of the single-skill Quantity Discrimination test.

Reliability averages of the two multi-skill screeners were somewhat lower than what had been found previously for the single-skill screener (i.e., .78 and .86 for the fall and spring administrations, respectively), but these reliability estimates fall within an acceptable range

(Urbina, 2004). In terms of concurrent and predictive validity, however, figures for the multi-skill screeners generally surpassed those of the single-skill screener. For example, with respect to fall-of-kindergarten to end-of-first-grade predictive validity, coefficients ranged from .55 to .72 for the two multi-skill math screeners with outcome measures (i.e., vs. .52 to .66 for the single-skill screener). 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 multi-skill screeners are higher than the average predictive validity of the kindergarten screening literature we reviewed (i.e., Baker et al., 2002; Bramlett, Rowell, & Mandenberg, 2000; Chard et al.; Clarke et al.; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Kurdek & Sinclair, 2001; Lembke & Foegen; Mazzocco & Thompson, 2005; Pedrotty Bryant et al.; Tiesl, Mazzocco, & Myers, 2001; VanDerHeyden, Witt, Naquin, & Noell, 2001), which comprises an assortment of screening and outcome measures. Kindergarten math screeners from these earlier studies correlated (on average) .46 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 evens out. 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--a harbinger of the resulting overall accuracy of the screeners in predicting MD.

Although it was not the sole focus of the present study, documenting the technical adequacy of the math screening assessments constituted an essential first step toward drawing conclusions about the screeners' predictive utility. Practically speaking, if educators and

diagnosticians are to rely on a test to forecast future MD status, the test must demonstrate reasonable levels of score stability and consistency. Furthermore, inferences drawn from the test's scores must be meaningful and justifiable (in this case, with respect to students' early mathematics ability). As did previous evaluations of the Quantity Discrimination measure (e.g., Chard et al.), our data lend support to its technical adequacy as well as that of the multi-skill Computation Fluency and Number Sense screeners.

In addition to examining the kindergarten math screeners, however, we were particularly interested in aspects of the screeners' decision-making utility. Only a handful of previous kindergarten screening studies looked beyond predictive validity correlations and directly analyzed the sensitivity or specificity of their screeners (Bramlett et al., 2000; Mazzocco & Thompson, 2005; Simner, 1982; Tiesl et al., 2001; VanDerHeyden et al., 2001). With the present study, we specifically questioned whether the predictive utility of our tests would differ 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. operational). To our knowledge, no previous work has addressed these concerns. If educators are to accurately pinpoint students in need of intensive math intervention (i.e., in an attempt to prevent future MD), research should inform the practice of *how*, *when*, and with respect to *what outcome* this may best be accomplished.

First, 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? Gersten et al. (2005) suggested that measures comprising items of counting/simple computation skill and quantity/use of mental number line may effectively screen young students for potential MD. Along these lines, we questioned whether a single aspect of "number sense" (i.e., such as quantity discrimination)

would prove sufficient as a predictor of MD. Alternately, to maximize effectiveness, we asked whether a screener comprising items of multiple early numeracy concepts would provide enhanced decision-making utility. To answer these questions, we compared the AUCs of the single-skill to the multiple-skill screeners, at both the fall and spring testing occasions, and with respect to two mathematical outcomes. Non-overlapping confidence intervals would indicate statistical differences between models.

Our results showed no significant differences in predictive utility for single- versus multi-skill screening, at fall or spring, for either math outcome. This is interesting, given that the predictive validity of the multi-skill screeners was generally higher than that of the single-skill screener. This highlights the importance of looking at the predictive utility of screening measures in addition to the simple predictive correlations. Our results indicate that a brief, timed 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 future MD. This is likely welcome news for kindergarten teachers who often have limited time and/or resources available to screen their classes of young learners. As a reminder, the single-skill quantity discrimination was a one-minute, timed probe; the multi-skill Computation Fluency screener was a 5-minute timed, group-administered test; and the multi-skill Number Sense test was untimed, individually administered, and took from 10 to 15 minutes per student to complete. Of course, separate from the issue of efficiency, the multi-skill screeners may provide teachers with better information for instructional planning than the single-skill screener. This is because sampling a wider variety of early mathematical skills, as the multi-skill screeners do, provide an opportunity for error analysis and for highlighting students' specific numerical strengths and weaknesses. The single-skill screener, on the other hand, provides information on only one aspect of mathematical skill.

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 is important to know, 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; Johnson, Jenkins, Petscher, & Catts, 2008), stressing the school system 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 initial low performance results from developmental or experiential lag rather than true MD. If this were the case, one would expect to uncover a significant difference in predictive accuracy from the fall to the spring testing occasions. On the other hand, refraining from screening students for MD until the spring of kindergarten (or even later), with the belief that fall screening is not trustworthy, denies students of months of intervention time that could well serve to offset or prevent extreme 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 statistical differences in predictive utility from the fall to the spring testing occasions, underscoring the potential value of beginning early, in the fall of students' kindergarten year, to identify young learners in need of mathematical intervention. In spite of this, the large numbers of false positives (i.e., ranging from 30 to 48 and from 41 to 61 for conceptual and operational outcomes, respectively) suggest that delaying screening until after kindergarten 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 and diagnosticians consider conceptual mathematical deficits as a hallmark of MD at the end of first grade, or conversely, should the focus be on operational deficits? Prior work shows that elementary-aged students with MD show marked deficits in computational

fluency and difficulty with number processing (e.g., Jordan et al. 2003; Mazzocco, 2007). Yet, it is plausible that students as young as 5 and 6 years may simply not have had sufficient or comparable formal instruction with paper-and-pencil tasks such as counting, addition, or subtraction facts. As such, choosing a math outcome to designate MD which focuses on operational skill for students at this young age (i.e., such as written number combinations or 2-digit addition and subtraction items) may prove less useful than one that focuses on early numeracy concepts more likely to have been taught with early math curricula (i.e., such as shape identification or the meaning of “more than” or “less than”). Our results supported this. When we contrasted predictive models with conceptual versus operational mathematical outcomes, we found those with conceptual outcomes to be statistically better than those with operational outcomes, regardless of type of screener (i.e., single- or multi-skill) or time of testing (i.e., fall or spring). During the fall or spring of kindergarten, AUCs for our screening models ranged from .80 to .88, indicating “good” predictive utility for conceptual outcome using the EMDA Math Reasoning subtest. By contrast, during the same time frames, AUCs ranged only from .64 to .72, indicating “poor” predictive utility for operational outcome using the EMDA Numerical Operations subtest. This suggests that we can predict future computational deficits less accurately than conceptual deficits, at least when screening learners in the kindergarten year.

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. Yet, with respect to operational outcome at the same time, the single- and multi-skill screeners produced similar but significantly less accurate fits. Although our results lend tentative support to the potentiality of screening students as young as kindergarteners for future MD, additional study is needed to increase the overall accuracy of this task. That is, regardless of the predictive model used, we found an unacceptably high

proportion of students misidentified as false positives and/or false negatives. 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; Johnson, 2008). This suggests the potential need for a multiple-gating screening procedure, in which a cut-point on the universal screen is set to minimize false negatives, and then a more thorough conventional assessment or a dynamic assessment or short-term progress monitoring is conducted among the subset of students who failed the universal screen. In reading, Compton, Fuchs, Fuchs, & Bryant (2006) showed how such a multiple-gating screening procedure, using six weeks of short-term progress monitoring at the beginning of the first grade could eliminate false positives and false negatives. Future work should investigate the potential of multiple-gating kindergarten screening procedures to identify risk of MD more precisely.

As readers interpret findings, however, at least four limitations to the study should be considered. Three pertain to the participants; one 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 of results. Second, although our attrition rate was within reason, 22%, it is unclear how results may have been affected had the 56 students who moved remained through the end of first grade. We however note that on the fall kindergarten multi-skill Number Sense screening measure, students who remained through the end of first grade scored significantly higher than those who exited. This finding raises questions about whether results would change if the exiters had remained. Even so, the students who exited and those who remained were demographically comparable. Moreover, 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. Finally, we did not address the issue of timed testing in this study. The single-skill quantity discrimination screener and the multi-skill Computation Fluency screener were timed; the multi-skill Number Sense screener was untimed. 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 drawing distinctions among students' skill levels, abilities, and potential. In any case, we cannot state whether timed tests makes a difference in predictive utility for students at this age.

To address these limitations, future research should employ a more representative sample and should systematically vary timed versus untimed administration of screening measures. Additionally, future research should evaluate how the use of our multi-skill measures for progress monitoring might enhance teachers' instructional planning and student learning. Finally, and in a related way, the role of multiple-gating screening processes should be investigated as a means of lowering the rate of false positives and false negatives.

Appendix A – Computation Fluency

COMPUTATION FLUENCY

Score: ___/25

Form A

Name: _____

Date: _____

$2 + 3 = \underline{\quad}$	$\begin{array}{cc} * & * \\ * & * \\ \hline \end{array}$	$\begin{array}{ccc} * & * & + & * & = \\ \hline \end{array}$	$4 - 2 = \underline{\quad}$	<p>Cross out 2 *.</p> $\begin{array}{cccc} * & * & * & * \\ \hline \end{array}$
$3 - 1 = \underline{\quad}$	<p>Cross out 4 *.</p> $\begin{array}{ccccc} * & * & * & * & * \\ * & * & * & & \\ \hline \end{array}$	$\begin{array}{ccc} * & * & \\ * & * & * \\ \hline \end{array}$	$\begin{array}{ccc} *** & + & ** & = \\ \hline \end{array}$	$0 + 4 = \underline{\quad}$
$\begin{array}{c} * \\ * \\ \hline \end{array}$	$2 + 2 = \underline{\quad}$	$5 - 1 = \underline{\quad}$	<p>Cross out 1 *.</p> $\begin{array}{ccc} * & * & * \\ \hline \end{array}$	$\begin{array}{ccc} ***** & + & **** & = \\ \hline \end{array}$
$\begin{array}{ccc} * & + & * * * * * & = \\ \hline \end{array}$	$3 - 3 = \underline{\quad}$	<p>Cross out 3 *.</p> $\begin{array}{ccccc} * & * & * & * & * \\ * & * & * & * & * \\ \hline \end{array}$	$1 + 4 = \underline{\quad}$	$\begin{array}{ccc} * & * & * \\ * & * & * \\ * & * & * \\ \hline \end{array}$
<p>Cross out 0 *.</p> $\begin{array}{ccccc} * & * & * & * & * \\ \hline \end{array}$	$\begin{array}{ccc} ***** & + & ***** & = \\ \hline \end{array}$	$3 + 1 = \underline{\quad}$	$\begin{array}{c} * \\ \hline \end{array}$	$5 - 3 = \underline{\quad}$

Appendix B – Number Sense

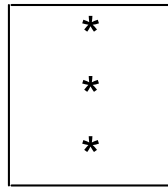
Score: ____/30
Age: _____

NUMBER SENSE

Name: _____

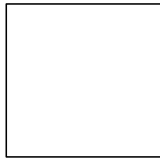
Date: _____

1)



4 19

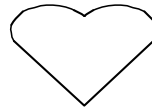
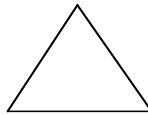
2)



3)

22	2	22	2	
-----------	----------	-----------	----------	--

4)



5)

6)

* * * * *
* * * * *

7)

2	8
---	---

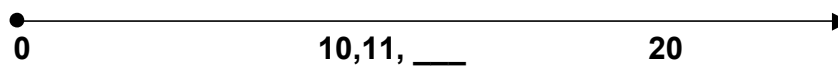
8)

4	0
---	---

9)

+	+	+	+	+	+	+	+	
---	---	---	---	---	---	---	---	--

10)



11)

12)

4, 3, 2, _____

13)

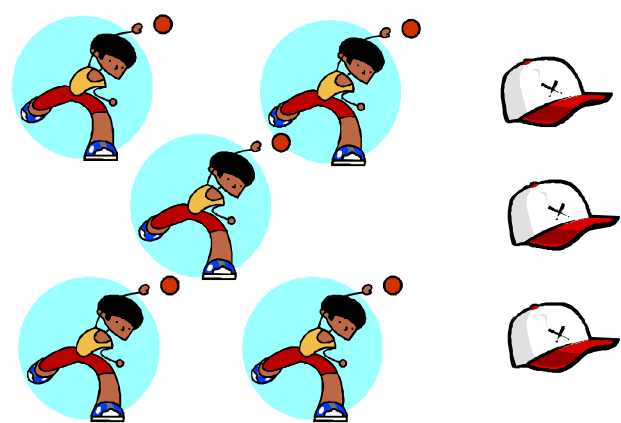
7	2
---	---

14) 2 4 3 _____

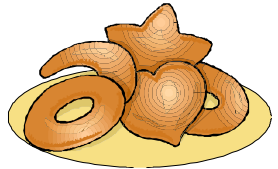
15)

17	16
----	----

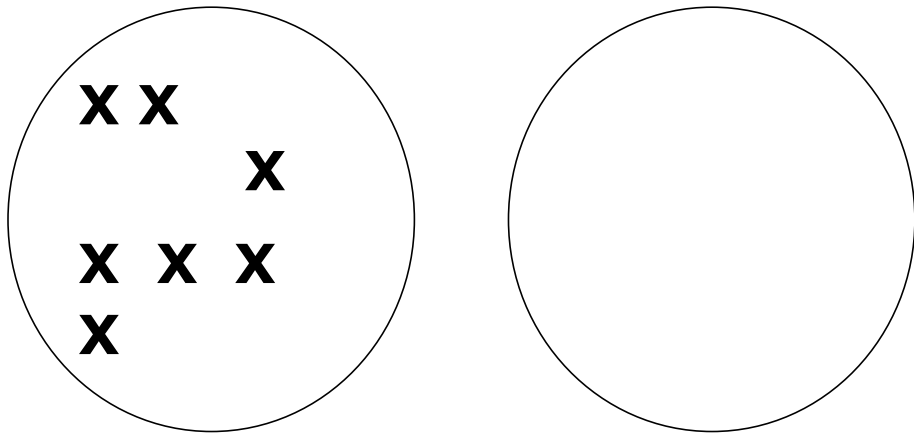
16) _____ boys



The illustration shows five boys in a circle, each with a red ball above their head. To the right of the boys are three baseball caps, each with a red brim and a white top with a red 'X' on the front.

17) $5 + 4 = 9$  $5 - 4 = 1$

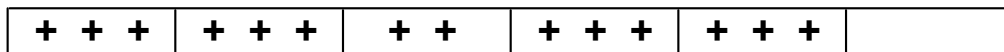
18)



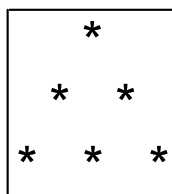
The first circle contains the following arrangement of 'X' marks:
X X
 X
X X X
X

The second circle is empty.

19)



20)



8

5

21)

9, 8, _____

22)

15

16

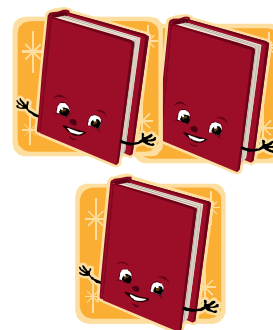
14



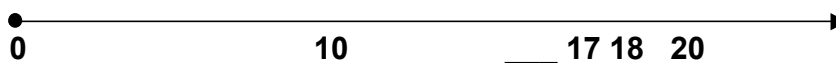
23)

$2 - 1 = 1$

$2 + 1 = 3$

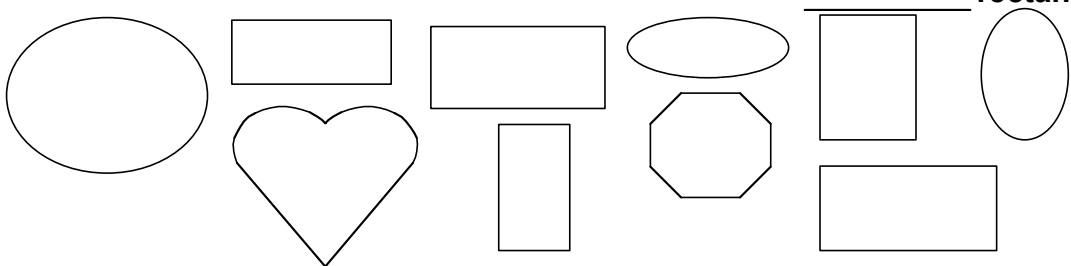


24)

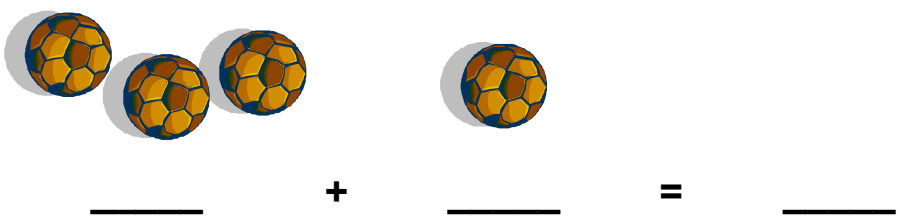


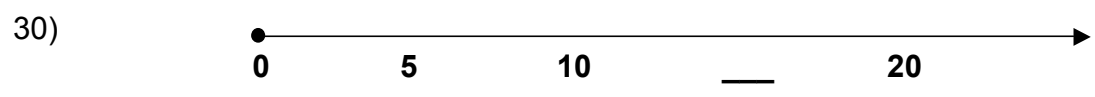
25) **20** **19** **18** _____

26) _____

27)  **rectangles**

28) **17,** _____

29)  _____ + _____ = _____



Appendix C – Number Sense Score Sheet

Number Sense Score Sheet

Now we're going to do some math activities.

Scoring

1 = correct response

0 = incorrect response

Ceiling = 5 (Stop after 5 consecutive scores of 0)

		Quantity Discrimination	Mental Number Line	Ordering Numbers	Estimation	Patterns	Counting Backward	Shape Discrimination	Number Sentences	Writing Numbers	One-to-one correspondence
1.	4										
2.	square										
3.	22										
4.	(marks triangle)										
5.	1,2,3,4,5										
6.	18										
7.	8										
8.	0										
9.	+++										
10.	12										
11.	6,7,8,9,10										
12.	1										
13.	2										
14.	2,3,4										
15.	17										
16.	2										
17.	(circles "5-4=1")										
18.	(draws 7 Xs)										
19.	++										
20.	5										
21.	7										
22.	14,15,16										
23.	(circles "2+1=3")										
24.	16										
25.	18,19,20										
26.	10,20,30,40,50										
27.	5										
28.	16										
29.	3+1=4										
30.	15										
Domain Scores											

Ceiling Item

Raw Score

REFERENCES

- Achenbach, T. M. (1991). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington: University of Vermont, Department of Psychiatry.
- Augustyniak, K. M., Cook-Cottone, C. P., & Calabrese, N. (2004). The predictive validity of the Phelps Kindergarten Readiness Scale. *Psychology in the Schools, 41*, 509-516.
- 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.
- Berninger, V. W., Thalberg, S. P., DeBruyn, I., & Smith, R. (1987). Preventing reading disabilities by assessing and remediating phonemic skills. *School Psychology Review, 16*, 554-565.
- 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.
- Brigance, A. (1999). *Comprehensive Inventory of Basic Skills (rev. ed.)*. North Billerica, MA: Curriculum Associates, Inc.
- Catts, H. (1991). Early identification of dyslexia: Evidence from a follow-up study of speech-language impaired children. *Annals of Dyslexia, 41*, 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.
- Compton, D., L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, (98)*, 394-409.
- Conners, C. K. (1997). *Conners' Rating Scales-Revised: Technical Manual*. Toronto, Ontario,

Canada: Multi-Health Systems.

- Connolly, A. J. (1998). *KeyMath-Revised*. Circle Pines, MN: American Guidance Service, Inc.
- Costenbader, V., Rohrer, A. M., & Difonzo, N. (2000). Kindergarten screening: A survey of current practice. *Psychology in the Schools, 37*, 323-332.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.
- 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.
- Ferri, B. A., & Connor, D. J. (2005). In the shadow of brown: Special education and overrepresentation of students of color. *Remedial and Special Education, 26*, 93-100.
- Fuchs, L. S., & Fuchs, D. (2004). *Using CBM for progress monitoring in math*. Available on www.studentprogress.org.
- 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., et al. (1994). Classwide curriculum-based measurement: Helping general educators meet the challenge of student diversity. *Exceptional Children, 60*, 518-537.
- Fuchs, L.S., Hamlett, C., & Fuchs, D. (1989; 1990). *Monitoring basic skills progress: Basic math*. For information, contact L. S. Fuchs, 228 Peabody College, Nashville, TN 37203.
- Gall, M. D., Gall, J. P., & Borg, W. J. (2003). *Educational research: An introduction* (7th Ed.). Boston, MA: Pearson Education, Inc.
- Geary D. C. (2003). Learning disabilities in arithmetic: Problem-solving differences and cognitive deficits. In K. R. Harris & H. L. Swanson (Eds.), *Handbook of learning disabilities* (pp. 199-212). New York, NY: Guilford.
- Geary, D. C., & Brown, S. C. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in gifted, normal, and mathematically disabled children. *Developmental Psychology, 27*, 398-406.
- Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology, 88*, 121-151.

- Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: Patterns of functions and deficits in children at risk for a mathematical disability. *Journal of Experimental Child Psychology*, 74, 213-239.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education*, 33, 18-28.
- 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.
- Hoover, H. D., Hieronymous, A. N., Dunbar, S. B., & Frisbie, D. A. (1993). *Iowa Test of Basic Skills*. Itasca, IL: Riverside.
- Individuals with Disabilities Education Act Amendments of 1997, Sec. 602(26), p. 13.
- Individuals with Disabilities Education Improvement Act (2004). Sec. 614(b)(6).
- 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.
- Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2008). *How can we improve the accuracy of screening instruments?* Manuscript under review.
- 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.
- Kaminski, R. A., & Good, R. H. III. (1996). Toward a technology for assessing basic literacy skills. *School Psychology Review*, 25, 215-227.
- Kelly, M. S., & Peverly, S. T. (1992). Identifying bright kindergartners at risk for learning difficulties: Predictive validity of a kindergarten screening tool. *Journal of School Psychology*, 30, 245-258.
- 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.

- 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.
- 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?* Baltimore, MD: Paul H. Brookes Publishing Co.
- Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice*, 20, 142-155.
- Mercer, C. D., Jordan, L., Allsopp, D. H., & Mercer, A. R. (1996). Learning disabilities definitions and criteria used by state education departments. *Learning Disability Quarterly*, 19, 217-232.
- National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.
- 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 Publishers.
- 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.
- President's Commission on Excellence in Special Education (2001). *A new era: Revitalizing special education for children and their families*. Washington, DC: US Department of Education.
- 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.
- Salvia, J., & Ysseldyke, J. E. (1991). *Assessment (5th Ed.)*. Boston: Houghton Mifflin Company.
- Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal member, rapid naming, and IQ. *Annals of Dyslexia, 48*, 115-136.
- Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- 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.
- Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Guide for administering and scoring the fourth edition Stanford-Binet intelligence scale*. Chicago: Riverside.
- 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. (1998). Catch them before they fall. Identification and assessment to prevent reading failure in young children. *American Educator, 22*, 32-39.
- Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: John Wiley & Sons, Inc.
- Valencia, R. R., & Suzuki, L. A. (2001). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues*. Thousand Oaks, CA: Sage Publications, Inc.
- VanDerHeyden, A. M., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27*, 27-41.
- 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.