

# A Preliminary Investigation Into the Identification and Development of Early Mathematics Curriculum-Based Measurement

Ben Clarke

*Pacific Institutes for Research, Eugene, Oregon*

Mark R. Shinn

*National-Louis University, Evanston, Illinois*

*Abstract.* Recent studies indicate that students in the United States are not achieving sufficient mathematics skills to meet the demands required of them within and outside of school. Among the keys to preventing mathematics difficulties are to identify and intervene with those students who may be most at-risk for later failure, monitoring their progress as frequently as possible. Unfortunately, current mathematics tests do not meet both these keys until mathematics instruction is well underway. This study examines the reliability, validity, and sensitivity of four experimental early mathematics measures designed for use in early identification and formative evaluation. The measures were based on the principle of number sense and were designed to assess the precursors of mathematics understanding learned before children are able to do formal mathematics. First grade students ( $N = 52$ ) were tested and interscorer, alternate form, test-retest reliability, and concurrent and predictive validity with three criterion measures were examined. Results showed that the four experimental measures each had sufficient evidence of their reliability, validity, and sensitivity. The differences in the utility of each experimental measure are analyzed from an early identification and formative evaluation perspective. Implications for practice are discussed.

A primary goal of instruction for schools is the development of students with mathematics skills. Mathematics is defined as a language that is used to express relations between and among objects, events, and times. The language of mathematics employs a set of symbols and rules to express these relations (Howell, Fox, & Morehead, 1993).

Proficiency in the language of mathematics is becoming an increasingly vital skill for all individuals in today's society. More than

10 years ago, the United States Department of Labor (1990) recognized the growing emphasis placed on technology in the marketplace. The demands of a new marketplace required greater proficiency by employees in mathematics. In this new environment, many of the fields projected to have the highest rate of growth in available jobs would be open only to individuals who are proficient in mathematics (United States Department of Labor, Bureau of Labor Statistics, 1997). The increased need for pro-

---

This research was supported in part by Grant No. 84.029D60057 Leadership Training in Curriculum-Based Measurement and Its Use in a Problem-Solving Model sponsored by the U.S. Department of Education, Office of Special Education. The views expressed within this paper are not necessarily those of the USDE. Address all correspondence and questions about this article to Ben Clarke, Pacific Institutes for Research, 1600 Millrace Dr., Suite 111, Eugene, OR 97403; E-mail: clarkeb@oregon.uoregon.edu

Copyright 2004 by the National Association of School Psychologists, ISSN 0279-6015

iciency in mathematics is exemplified further by the hiring practices of companies. Companies are often requiring workers to have a minimum of mathematics skill, even if the job is one that is not typically associated with the need for skill in this area. Once employed, individuals who are proficient in mathematics earn 38% more than individuals who are not proficient in mathematics (Riley, 1997).

Given the pressing need for mathematics proficiency, an examination of the current state of mathematics performance in the United States is warranted. Recent national studies indicated that the current performance of United States students may be such that students will not have the necessary skills to meet the changing demands of the United States workplace. The 1996 National Assessment of Educational Performance (Reese, Miller, Mazzeo, & Dossey, 1997) results classified only 21% of fourth-grade students as at or above proficiency in mathematics performance. The low level of students classified as having proficient skills stands in sharp contrast to the classification of 36% of fourth-grade students as below basic in mathematics performance. The pattern of large numbers of students scoring below the basic level and few students scoring at or above proficient level is also found in Grades 8 and 12.

### **Assessment Solutions to Mathematics Problems**

The demand for students skilled in mathematics coupled with current low levels of achievement suggest a need to examine ways to increase achievement. A number of critical variables have been identified that are related to general increases in student achievement (Brophy & Good, 1986). Foremost among these variables are (a) early intervention and (b) formative evaluation of student progress.

Early intervention can be defined as “formal attempts by agents outside the family to maintain or improve the quality of life of youngsters starting with the prenatal period and continuing through entry into school” (Karoly et al., 1998, p. 4). To maximize the effectiveness of early intervention, individuals without critical skills or who are at risk for failure to

develop these skills must be identified as soon as possible. Additionally, formative evaluation needs to be incorporated into early intervention as a way of measuring growth and improvement. Formative evaluation is a methodology of testing during the instructional process for the purpose of evaluating the effectiveness of an intervention and monitoring student performance and growth (Deno, 1985, 1986). The information gathered in formative evaluation improves outcomes by providing an empirical database that can be used to design and modify instructional programs to be more effective (Fuchs, 1986). A meta-analysis by Fuchs and Fuchs (1986) found that when instructional programs employ systematic formative evaluation and use the information to develop and monitor programs over time, student achievement increased an average of .7 standard deviation units.

Unlike more summative assessment (i.e., one-shot achievement tests), formative evaluation measures must be designed differently according to features identified by Fuchs and Fuchs (1999). Formative evaluation measures must be evaluated by the following criteria: (a) technical adequacy, (b) capacity to model growth, (c) treatment sensitivity, (d) instructionally eclectic, and (e) feasibility.

### **Developing Assessment Tools for Mathematics Early Identification and Formative Evaluation**

Once students have begun explicit instruction in mathematics it is possible to identify at-risk students and monitor achievement interventions using mathematics curriculum-based measurement (M-CBM). M-CBM, like other CBM measures, is based on a validated, standard, simple to administer and score, short-duration fluency measure where students write answers to computational problems for 2 minutes.

One obvious problem with M-CBM is the difficulty in using M-CBM in early identification. M-CBM measures, like other math achievement tests, are reactive to identifying problems in mathematics achievement. M-CBM can be used once students are expected to be learning mathematics and there is suffi-

cient mathematics growth to measure. For many students, this means that M-CBM can be used in mid to late first grade at best. Prior to this point, nearly all students will have initial scores of 0 on M-CBM probes and thus early identification is hampered by floor effects. The inability to use M-CBM in early identification may result in a failure of students to acquire critical mathematical concepts and skills related to subsequent math achievement.

### **Viable Mathematics Concepts for Use as Early Mathematics-CBM (EM-CBM)**

The acquisition of basic numerical concepts during early childhood serves as a foundation for the acquisition of later higher order mathematical concepts (Ginsburg & Allardice, 1984). Thus, failure to attain critical early numerical concepts can influence later ability to acquire important mathematic skills. A failure in understanding and acquiring early numerical concepts can also influence the level of interest and confidence a student brings to new experiences with numbers and math. Early failure in mathematics has the potential to fundamentally alter the child's mathematics education program throughout their time in school (Jordan, 1995).

One frequently mentioned outcome of informal early math learning is the beginning of *number sense development*. Although a formal definition of number sense has not been set forth, there is a growing consensus regarding what number sense is and what characteristics make up number sense. The National Council of Teachers of Mathematics (Commission on Standards for School Mathematics, 1989) described children with number sense as having the ability to understand the meaning of numbers and define different relationships among numbers. Children with number sense can recognize the relative size of numbers, use referents for measuring objects and events, and think and work with numbers in a flexible manner that treats numbers as a sensible system (Resnick, 1989). A potential working definition of number sense that encapsulates the critical features of number sense and its role in mathematics was offered by Gersten and Chard (1999) who defined number sense

as "a child's fluidity and flexibility with numbers, the sense of what numbers mean, and an ability to perform mental mathematics and to look at the world and make comparisons" (p. 20).

The need to develop number sense is reflected in Standard 6 of the Curriculum and Evaluation Standards for School Mathematics for students in kindergarten through fourth grade. One of the primary goals for students is to develop number sense (Commission on Standards for School Mathematics, 1989). In addition, the National Research Council (1989) proclaimed, "The major objective of elementary school mathematics should be to develop number sense" (p. 46). Based on research that suggested students entering kindergarten differed on their ability to complete basic activities requiring number sense, Griffin, Case, and Siegler (1994) suggested that early intervention efforts be focused on building and expanding the informal number sense that children bring to school. Subsequent research studies (Griffin et al., 1994) provided preliminary evidence that programs to build number sense were successful in providing students with number sense at the conclusion of the program and that the expanded number sense knowledge was still prevalent at later dates. In addition to the importance of developing instructional programs to build number sense, an important first step of mathematics research into number sense is the investigation of assessment tools that could be used to identify students at-risk for later failure in mathematics Gersten and Chard (1999).

The purpose of this study was to investigate four potential EM-CBM measures of number sense. An oral counting measure, a number identification measure, a quantity discrimination measure, and a missing number measure were examined as potential EM-CBM measures for use in early identification and formative evaluation. The measures were evaluated from a sequenced set of standards. First, the measures needed to meet reliability criteria. If reliability criterion were met, the validity of the measures from both a concurrent and predictive standpoint must be established. Finally, the measure's sensitivity to stu-

dent growth was examined as a necessary component for potential use in formative evaluation.

## Method

### Participants

Participants in this study were 52 first-grade students from the Fall 2000 to Spring 2001 academic year. Participants were recruited through a letter sent home to their parents explaining the purpose of the study. Parents returned the letter indicating whether or not they wished their child to participate. Parents who did not return the letter were considered to have not granted permission for participation. All participants were receiving their schools' standard first-grade mathematics curriculum. Participants were drawn from two schools in a medium-size school district of 2,500 students located in the Pacific Northwest. The two schools had 58% and 59% of their students qualified for free and reduced lunch. Of the 52 participants, 29 were female and 23 were male. The majority of the participants were Caucasian, with 2 students who were Native American and 3 who were Hispanic. Three of the participants (6%) were receiving special education services.

### Measures

Participants were administered seven different measures of mathematics over the duration of the study, four experimental measures and three criterion measures. The four experimental measures were Oral Counting (OC), Number Identification (NI), Quantity Discrimination (QD), and Missing Number (MN). The criterion measures were the Woodcock-Johnson Applied Problems subtest (WJ-AP; Woodcock & Johnson, 1989), the Number Knowledge Test (NKT; Griffin, Case, and Siegler, 1994), and math curriculum-based measurement (M-CBM) first-grade computation probes.

### Experimental Measures

Each of the experimental measures was individually administered. All were 1 minute in duration. Copies of the standard directions

and sample measures are available from the first author.

**Oral Counting Measure (OC).** The OC measure required students to count orally, starting with 1. No student materials were used. An examiner recorded student responses by following along on a scoring sheet. Numbers that were correctly counted in sequence were scored as correct. Numbers that were not correctly counted in sequence were scored as incorrect. If a participant stopped, struggled, or hesitated to say a number for 3 seconds, they were instructed to say the next number. Participant performance on the OC measure was reported as the number of numbers correctly counted in 1 minute.

**Number Identification Measure (NI).** The NI measure required participants to identify orally numbers between 0 and 20 when presented with a set of printed number symbols. Participants were given a sheet of randomly selected numbers formatted in an 8 by 7 grid. Numbers that were correctly identified were counted as correct. Numbers that were not correctly identified or skipped were counted as incorrect. If a participant struggled or hesitated to correctly identify a number for 3 seconds, they were instructed to "try the next one." Participant performance was reported as the number of numbers correctly identified in 1 minute.

**Quantity Discrimination Measure (QD).** The QD measure required participants to name which of two visually presented numbers was larger. Participants were given a sheet of paper with a grid of 32 individual boxes. In each box were two randomly sampled numbers from 0 to 20. In the box, one number was always larger than the other number. Boxes in which the participant correctly identified the larger number were counted as correct. Boxes in which the participant named the smaller number, named any number other than the bigger number, or did not state a number were counted as incorrect. If a participant stopped, struggled, or hesitated for 3 seconds, the participant was instructed to "try the next one." Participant performance was reported as the number of correctly identified larger numbers in 1 minute.

**Missing Number Measure (MN).** The MN measure required the participant to name the missing number from a string of numbers between 0 and 20. Students were given a sheet with 21 boxes on it. In the boxes were strings of three numbers with either the first, middle, or last number of the string missing. The participant was instructed to orally state the number that was missing. Numbers missing that were correctly identified were counted as correct. Numbers missing that were not correctly identified or skipped were counted as incorrect. If a participant struggled or hesitated to correctly identify a number missing for 3 seconds, they were provided the number by the examiner and instructed to "try the next one." Participant performance was reported as the number of missing numbers correctly identified in 1 minute.

### Criterion Measures

Criterion measures were utilized to help examine the concurrent and predictive validity of the experimental measures.

**Math CBM Grade 1 Computation Probes (M-CBM).** M-CBM Grade 1 computation probes required the participant to write answers to math problems drawn from students' first-grade mathematics curriculum for 2 minutes. M-CBM probes were individually administered. Math probes were composed of addition and subtraction problems up to two digits without regrouping. Standardized M-CBM scoring and administration procedures were followed (Shinn, 1989). Each participant completed three first-grade M-CBM probes. For each probe, the examiner scored the number of correct digits with an answer key and the median score from the three probes was used in later analysis.

M-CBM interscorer agreements are reported as .93 to .98 (Marston, 1989). The internal consistency and test-retest reliability of M-CBM probes were reported as .93 for both types (Fuchs, Fuchs, & Hamlett, 1988; Tindal, Germann, & Deno, 1983). Correlations with the Metropolitan Achievement Test (MAT) Problem Solving and Math Operations for M-CBM are reported to range from .26 to .65. For younger

students, the correlations are lower. At the first- and second-grade levels, correlations have been reported of .26 to .34 with the MAT (Skiba, Magnusson, Marston, & Erickson, 1986).

**WJ-R Applied Problems subtest (WJ-AP; Woodcock & Johnson, 1989).** The WJ-AP subtest is an individually administered norm-referenced test of applied mathematics consisting of problems that require the use of mathematics operations to solve a variety of applied math problems. For students in first grade, the use of addition and subtraction operations are required to solve problems.

The WJ-AP subtest has a reported internal-consistency reliability coefficient of .84 for the 6-year-old norming group. Validity data for the WJ-AP subtest were collected as part of a larger examination of the validity of the WJ-R Broad Math Cluster. Concurrent validity evidence for the WJ-R Broad Math Cluster was collected with groups of children ages 3, 9, and 17. The age 9 group most closely approximates the estimated age of the participants in this study. Correlations ranged from .41 to .83 with a median of .71 with five criterion measures of mathematics including the BASIS-Math, KABC-Arithmetic, and KTEA-Math Composite among others.

**Number Knowledge Test (NKT; Okamoto & Case, 1996).** The NKT test contains four levels and students are required to obtain a minimum number of correct responses at one level to move to the next level. On Level 1, students are required to complete tasks such as counting chips and geometric shapes. Level 2 requires students to do tasks such as identifying bigger or smaller numbers from a pair, naming numbers, and solving simple addition and subtraction problems. Level 3 requires students to solve problems similar to those of Level 2, but with larger numbers. Level 3 also requires students to complete new items such as stating how many numbers are between a pair of numbers. Level 4 is a more difficult version of Level 3 and also adds new tasks such as telling which difference between two pairs of numbers is bigger or smaller.

Limited evidence for technical adequacy of the NKT was found. Evidence consisted of

**Table 1**  
**Data Collection Schedule**

Measures	Fall (00)	Winter (01)	Spring (01)
<b>Experimental</b>			
Oral Counting	1 probe	2 probes	1 probe
Number Identification	2 probes	2 probes	1 probe
Quantity Discrimination	2 probes	2 probes	1 probe
Missing Number	2 probes	2 probes	1 probe
<b>Criterion</b>			
Woodcock-Johnson Applied Problems	1 subtest		1 subtest
Number Knowledge Test	1 measure		
Mathematics Curriculum- Based Measurement		3 probes	3 probes

its use to evaluate an early mathematics curriculum (Griffin et al., 1994). The NKT was sensitive to student improvement and differentiated between participants in the treatment and control groups.

### Data Collection

Data were collected three times during the 2000-2001 academic year in the months of October (Fall), February (Winter), and May (Spring) with approximately 13 weeks between data collection periods. The data collection schedule detailing the timing of the administration of all is displayed in Table 1. Administration of experimental measures was counterbalanced. Per participant, the Fall data collection took approximately 35 minutes, Winter data collection took approximately 20 minutes, and Spring data collection took 25 minutes.

Examiners with a background in early childhood assessment were trained to administer and score the EM-CBM measures, M-CBM probes, the NKT, and the WJ-AP during a 2-hour training session. The training session consisted of instruction on administering the measures according to standardized directions and following standard protocols for scoring

the measures. Data collectors were observed administering and scoring each measure and appropriate feedback was provided. Feedback included what to do in cases where the participant failed to supply an answer or when the participant skipped items.

## Results

### Descriptive Statistics

Means and standard deviations for all measures in the Fall, Winter, and Spring are reported in Table 2. An examination of Table 2 indicates two patterns. First, across all four experimental measures, participants' scores improved implying change over time. Second, across each experimental measure, participants scored highest on the Oral Counting (OC) measure followed by the Number Identification (NI) measure, the Quantity Discrimination (QD) measure, and the Missing Number (MN) measure. Thus, the experimental measures had a consistent rank ordering at each point of the data collection. The standard deviations for each of the EM-CBM measures were consistent at each point of the data collection and followed the same pattern of rank ordering.

**Table 2**  
**Descriptive Statistics for First-Graders' Fall, Winter, and Spring Performance**  
**on EM-CBM and Criterion Measures ( $N = 52$ )**

Measures	Fall		Winter		Spring	
	<i>M</i>	( <i>SD</i> )	<i>M</i>	( <i>SD</i> )	<i>M</i>	( <i>SD</i> )
<b>Experimental</b>						
Oral Counting	60.4	(20.5)	69.7	(23.1)	74.6	(20.9)
Number Identification	36.0	(15.9)	42.5	(17.3)	48.1	(17.8)
Quantity Discrimination	19.2	(10.6)	23.9	(10.9)	28.5	(9.9)
Missing Number	11.3	(5.8)	14.5	(6.3)	17.4	(6.8)
<b>Criterion</b>						
Number Knowledge Test	12.2	(4.5)	—		—	
Woodcock-Johnson Applied Problems	104.5	(16.3)	—		107.7	(20.7)
Math-CBM	—		9.5	(7.0)	12.1	(7.9)

*Note.* All scores are reported in raw score units except for Woodcock-Johnson Applied Problems which is a standard score ( $M = 100$ ,  $SD = 15$ ). Scores for Number Identification, Quantity Discrimination, and Missing Number in the Fall are the average score of Form 1 and Form 2. Scores for Oral Counting, Number Identification, Quantity Discrimination, and Missing Number in the Winter are the average score of Form 1 and Form 2.

Based on the WJ-AP, participants in this study were in the average range compared to a national normative sample. In the Fall, participants scored 104.5 or at the 62nd percentile. In the Spring, participants scored 107.7 or the 69th percentile.

### Research Question 1: Reliability

The interscorer, alternate-form, and test-retest reliability of the experimental early math measures are reported in Table 3. Reliabilities were computed and analyzed using Pearson product moment correlation coefficients. Salvia and Ysseldyke (1998) provide a set of criteria with which to evaluate reliability in the context of educational decision making. Reliabilities of .90 or greater are recommended for making educational decisions about individual students. Reliabilities of .80 or greater are recommended for making screening decisions about individuals, and reliabilities of .60 or greater are recommended for making edu-

cational decisions about groups of students. Decisions made from early identification measures typically do not involve a high-stakes decision to change an individual student's placement or educational classification (e.g. eligibility for special education services; Kaminski & Good, 1998). Thus, although measures as free from measurement error as possible are desirable, the reliability standard of .80 was used as a minimum standard to evaluate reliability.

Interscorer reliability was assessed during the Fall data collection only, using 12 (23%) student protocols. Interscorer data were collected by having the lead author simultaneously score all the experimental measures for an entire testing session with a student. Each data collector utilized in the study was included in the sample to calculate interscorer reliability. Interscorer reliability was calculated as the number of items for which both data collectors agreed divided by the numbers of items for which both scorers agreed plus the number

**Table 3**  
**Early Mathematics Curriculum-Based Measurement (EM-CBM)**  
**Reliability For All Testing Sessions**

EM-CBM Measure	Inter-Scorer <sup>a</sup>	Alternate-Form		Test-Retest	
		Fall <sup>a</sup>	Winter <sup>b</sup>	13wks <sup>b</sup>	26wks <sup>b</sup>
Oral Counting	.99	—	.93	.80	.78
Number Identification	.99	.89	.93	.85	.76
Quantity Discrimination	.99	.93	.92	.85	.86
Missing Number	.98	.83	.78	.79	.81

*Note.* Test-retest reliabilities based on Form A.

<sup>a</sup>*n* = 12, <sup>b</sup>*n* = 52.

of items for which the scorers disagreed. Interscorer reliability for all experimental measures was high with .99 for the OC, NI, and QD measures and .98 for the MN measure, exceeding the standard for making individual educational decisions.

Alternate-form reliability was examined during the Fall and Winter data collection periods. These data are summarized in Table 3. In the Fall, students were tested on alternate forms for the NI, QD, and MN measures, but not the OC measure. Students were tested on alternate forms of all four measures during the Winter testing session. At this time, participants repeated the OC measure (i.e., counting from 1) because there was only one possible form of this measure. The order in which alternate forms for the NI, QD, and MN measures, or repeated for the OC measure, were given to participants was alternated to avoid practice effects.

Alternate-form reliability for the OC measure, NI measure, and the QD measure was consistently high and attained the .90 benchmark for individual educational decisions. Reliability for this MN measure was lower in the Fall (.83) and Winter (.78), but approximated the .80 standard for screening decisions.

Long-term test-retest reliability for all participants was examined from the Fall to Winter (13 weeks) and from the Fall to Spring

(26 weeks). All measures were acceptable, approaching or exceeding .80.

### **Research Question 2: Concurrent Validity**

Concurrent validity was assessed by examining correlations among the four experimental EM-CBM measures and the three criterion measures, the WJ-AP and the M-CBM and NST. These relations were examined at each of the three data collection periods. Concurrent criterion-related validity data were collected in the fall with the WJ-AP and the NST, in the winter with M-CBM, and in the spring with the WJ-AP and M-CBM.

Concurrent validity coefficients among the experimental measures and between the experimental and criterion measures by time frame are reported in Table 4. To provide a framework for interpreting the validity results, obtained correlations among the criterion measures (i.e., the NKT, WJ-AP subtest, and M-CBM probes) were examined. Concurrent validity correlations among the criterion measures ranged from .74 to .79 suggesting that the construct of first-grade mathematics was measured.

**Within experimental measures.** In general, the intercorrelations among all experimental measures were high. Notably, the NI,

**Table 4**  
**Concurrent Validity Correlations**

Measures	1	2	3	4	5	6
Fall						
1. OC	—					
2. NI	.69	—				
3. QD	.77	.93	—			
4. MN	.68	.84	.90	—		
5. NKT	.70	.70	.80	.74	—	
6. WJ-AP	.64	.65	.71	.68	.79	—
Winter						
1. OC	—					
2. NI	.71	—				
3. QD	.72	.88	—			
4. MN	.69	.82	.88	—		
5. M-CBM	.49	.66	.71	.75	—	
Spring						
1. OC	—					
2. NI	.68	—				
3. QD	.68	.86	—			
4. MN	.55	.72	.87	—		
5. M-CBM	.50	.60	.75	.74	—	
6. WJ-AP	.60	.63	.79	.69	.74	—

*Note.* OC = Oral Counting; NI = Number Identification; QD = Quantity Discrimination; NKT = Number Knowledge Test; WJ-AP = Woodcock-Johnson Applied Problems; M-CBM = Mathematics Curriculum-Based Measurement.

QD, and MN measures had consistently higher intercorrelations with each other than they did with the OC measure. Across all three testing periods, the intercorrelations for the NI measure ranged from .72 to .93 with a median of .85, for the QD measure from .86 to .93 with a median of .88, and for the MN measure from .72 to .90 with a median of .86 when their intercorrelation with OC was excluded. In contrast, the intercorrelations for the OC measure ranged from .55 to .79 with a median of .69.

#### **Relations with criterion measures.**

Relations among the experimental and criterion measures were then examined. Concurrent validity correlations were strongest for the QD measure ranging from .71 to .88 with a median of .75. Of the experimental measures, the OC measure had the lowest correlations ranging from .49 to .70 with a median of .60. The NI measure and MN measure concurrent validity evidence were between the two with the NI measure ranging from .60 to .70 with a

**Table 5**  
**Predictive Validity Correlations Between Experimental**  
**and Criterion Measures**

Measure	M-CBM	M-CBM	WJ-AP
	Winter	Spring	Spring
<b>Fall</b>			
Oral Counting	.56	.56	.72
Number Identification	.68	.60	.72
Quantity Discrimination	.76	.70	.79
Missing Number	.78	.67	.72
<b>Winter</b>			
Oral Counting	—	.46	.68
Number Identification	—	.58	.68
Quantity Discrimination	—	.71	.79
Missing Number	—	.72	.71

median of .66 and the MN measure ranging from .68 to .75 with a median of .71.

Further analysis of the concurrent validity correlations was done via a test for differences between two dependent correlation coefficients (Glass & Hopkins, 1996). Based on these comparisons, it was demonstrated that the QD measure was a better measure of early mathematics. The QD measure had significantly higher correlations than the OC measure with M-CBM in the Winter,  $t(49) = 2.93, p < .05$  and again with M-CBM in the Spring,  $t(49) = 3.31, p < .05$ . The QD measure also had significantly higher correlations than the NI measure with the NKT in the Fall,  $t(49) = 3.18, p < .05$ , with M-CBM in the Spring,  $t(49) = 3.03, p < .05$ , and with the WJ-AP in the Spring,  $t(49) = 3.50, p < .05$ . One other significant difference was found with the MN measure having a stronger relationship with M-CBM in the Winter,  $t(49) = 3.50, p < .05$  than the OC measure.

### Research Question 3: Predictive Validity

Predictive validity data were analyzed using the Fall EM-CBM measures and M-

CBM collected in the Winter and the WJ-AP and M-CBM data collected in the Spring. Predictive validity was also assessed between the Winter EM-CBM measures and the WJ-AP and M-CBM data collected in the Spring. The predictive validity of the experimental measures is summarized in Table 5.

The QD measure had the highest median correlation of .76, followed by the MN measure (.72). Both the NI measure (.68), and the OC measure (.56) had strong relationships as well.

Additional evidence further supported stronger predictive validity evidence for the QD and MN measures. Comparisons between the strength of correlations for the EM-CBM measures were conducted by testing the differences between two dependent correlation coefficients. The QD measure had significantly higher correlations than the OC measure with M-CBM from Fall to Winter,  $t(49) = 3.18, p < .05$ , and from Winter to Spring,  $t(49) = 3.34, p < .05$ . The MN measure was also more highly related to M-CBM than the OC measure from both the Fall to Winter,  $t(49) = 3.08, p < .05$ , and from Winter to Spring,  $t(49) = 3.34, p < .05$ .

#### Research Question 4: Sensitivity

Further evidence of validity is provided by examining the sensitivity of the EM-CBM measures over time. If the measures were assessing early mathematics achievement, then it was hypothesized that as participants progressed throughout the school year, they would learn more mathematics content and thus obtain higher scores on the EM-CBM measures.

Participants' scores improved on each of the four experimental measures across the 26-week period with students improving on the OC measure 14.2 units (.55 units per week), the NI measure 9.3 units (.47 units per week), the QD measure 9.3 units (.36 units per week), and the MN measure 6.1 units (.23 units per week). To further examine growth over time a repeated measures ANOVA was conducted. Growth for each measure was found not to be due to chance using the parameters of  $F(1,51), p < .01$ .

#### Discussion

The purpose of early intervention is to prevent severe problems from developing. To aid early intervention, assessment for this purpose must meet a number of important criteria. Like all assessment measures, those used in early identification should meet stringent criteria for reliability and validity. Educators who utilize measures in early identification should have confidence that the scores obtained are accurate scores across rater, form, and time and that the scores represent the critical concepts they are meant to measure. In addition to traditional criterion for measurement, measures used in early identification also must be able to identify those students most at-risk for failure and then be useful in monitoring their growth on critical skills over time. The results from this study indicate that the EM-CBM measures may be potentially useful in accomplishing these goals.

A number of significant findings emerge from the analysis of the reliability data collected. First, although there were differences in the strength of reliability data for each measure, no measure had poor reliability. In fact, the lowest reliability was .77. Second, the QD

measure was the most reliable experimental measure. All reliability coefficients exceeded the .80 criteria and four out of six values exceeded .90. The NI measure appeared to be the next most reliable experimental measure. The OC and MN measures each had strong evidence supporting their reliability, but were not as reliable as the QD and NI measures. All four EM-CBM measures met critical reliability criteria for making screening decisions about individual students.

The EM-CBM measures had strong evidence that they measure the skills they are intended to measure. Generally moderate to strong evidence of concurrent validity for all experimental measures was found. Median concurrent validity coefficients ranged from a low of .60 for the OC measure to a high of .75 for the QD measure. The predictive validity correlations between experimental and criterion measures showed relatively the same pattern as was found for concurrent criterion validity with the strongest evidence for the QD measure followed by the MN measure, the NI measure, and the OC measure. Further evidence that suggested the QD measure was the strongest measure of early mathematics was found by comparisons to other experimental measures that showed the QD measure as having significantly stronger relationships with the criterions. It is important to note that although the QD measure may have stronger evidence of measuring early mathematics, no measure had weak evidence.

The strength of reliability and validity evidence varied by measure. Across both reliability and validity, the QD measure had the strongest support for use as a single indicator of early mathematics. The OC measure had the weakest, albeit still strong, support. The NI and MN measures fell in between with the NI having greater evidence of reliability and the MN measure showing greater evidence of validity.

As a starting point for the development of a measure designed for early identification, evidence of reliability and validity are necessary but not sufficient. Due to the unique purposes of measures used in early identification, other factors are paramount as well. Measures

to be used in early identification must avoid floor effects or in simpler terms they must avoid multiple students showing little or no behavior on the measure. Although both the QD and MN measures were reliable and valid measures, their usefulness in early identification in the Fall may be less than the OC and NI measures because more students scored in a lower range (e.g., between 1–10) on these measures.

As stated above, reliability and validity are necessary but not sufficient evidence by which to evaluate measures designed to be used for early identification and formative evaluation. The amount of growth, or sensitivity, is important to examine in determining whether or not a measure might have potential to be used in the formative evaluation of a student's academic program. The results of repeated measures (F-test outcomes) indicated that each experimental measure accurately measured change over time and thus met a key criterion for use as a progress-monitoring measure. The OC measure appeared to be the most sensitive of the experimental measures followed by the NI, QD, and MN measures. The OC and NI measures may enable quicker decisions regarding student growth that could be seen by educators as true growth and not random fluctuation in scores. However, the measure that showed the greatest growth, OC, had the lowest reliability and validity coefficient correlations. The other measures had higher reliability and validity coefficients, but were not as sensitive to growth.

For example, the QD measure scores increased by about nine units over the 26-week course of the study. Thus, it would take between 2 and 3 weeks for a student to increase their score by one unit. Such a lack of sensitivity to student growth may make it difficult for educators to make timely decisions about educational progress. If the QD measure were to be used to evaluate a student's growth in early mathematics, it might be close to 2 months before educators were able to make a confident decision about progress. For an at-risk student, 2 months of failure may have an undesirable effect on their acquiring critical knowledge.

## Limitations and Future Research

As is the case with any research study, the conclusions drawn must be viewed within the context of the study's limitations. Foremost of the limitations is external validity. Participants were first-grade students from the Northwest United States. The generalizability of findings to other geographic areas, grades, and students should be investigated further.

External validity limitations are further compounded by the sample size of the study. Although sample size was similar to that found in other studies investigating curriculum-based measures for use in early intervention (Kaminski & Good, 1996; VanDerHeyden, Witt, Naquin, & Noell, 2001), the results of the study should be considered preliminary pending replication. In particular, further research into these measures should be done with a greater sample size to investigate interrater reliability and collect preliminary short-term test-retest data. Greater sample sizes would allow more definitive conclusions regarding the measures' potential and would allow more advanced statistical analysis to examine construct validity using factor analysis and growth over time using hierarchical linear modeling.

In addition, the preliminary nature of this study raises further questions to be investigated in subsequent research. Critical questions yet to be answered by this research fall under one of three broad categories. First, are there ways to improve the measures for first-grade students? Second, are these measures valid for younger students such as kindergartners? Third, how strong is the long-term predictive validity of these measures?

All the early math measures were reliable and valid measures of the early mathematics for first-grade students. However, an ongoing challenge in developing measures for use in early identification is examining the relationship between the reliability and validity of the measures and their potential use in early identification and formative evaluation. Future research should

attempt to further parcel out the relationship between reliability, validity, and the measures' utility in educational decision making.

Perhaps the best answer to which, if any, of the four measures "works best" will be answered by research with kindergarten students. The participants in this study were all taken from a common grade (first grade) rather than a cross-sectional design with participants also being drawn from kindergarten. Because of emphasis on early intervention as soon as possible, we cannot be sure that first grade is early enough to prevent the development of math problems. For example, although the experimental math measures were valid and reliable for first-grade students, would they show the same technical characteristics with younger students in kindergarten? A research study that examined the utility of the four experimental early math measures for kindergarten students would help to determine whether or not the early math measures were useful for early identification at any point in the kindergarten year. The measures as designed may prove to be too difficult for kindergarten students, thus resulting in multiple students who show little to no behavior on the measure. Such results may mean the measures would have limited utility in early identification and formative evaluation. One potential modification could be to limit the range of responses from 1 to 10 versus the current range of 1 to 20.

Extending research to kindergarten students also allows for an investigation of what type of assessment matrix is needed to track the growth of early math skills prior to the use of M-CBM in second grade. The use of multiple measures as an index of early mathematics skills would require more complexity and difficulty in implementing early mathematics assessment into everyday practice. This raises the question of whether there is a single indicator of early mathematics that can be used across years until the students are performing math and M-CBM can be used to measure growth. Alternatively, does early mathematics assessment require multiple measures to accurately assess critical early math skills?

The long-term predictive validity aspect of this study extended over the duration of 1 school year. Future research should attempt to extend the duration of time between initial measurement on the early math measures and criterion measures at later dates such as the end of second grade, third grade, and so on. A key aspect of the DIBELS assessment matrix is the ability to determine whether or not a student is on the right trajectory to becoming a reader if they have attained a specific criterion score on a DIBELS measure at an earlier date. In the same vein, what are critical criterion scores on the experimental math measures that allow a great enough degree of confidence to state that a student is or is not on the trajectory to becoming a fluent mathematician?

The results from this study provide a foundation upon which to further investigate the potential for the effective use of early mathematics measures to be used in early identification and formative evaluation. Continued research in this area should enable educators to provide better services to children attempting to acquire mathematics skills fundamental to their success in school and later in life.

## References

- Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328-375). New York: MacMillan.
- Commission on Standards for School Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review*, 15, 358-374.
- Fuchs, L. S. (1986). Monitoring progress among mildly handicapped pupils: Review of current practice and research. *Remedial and Special Education*, 7(5), 5-12.
- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199-208.
- Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review*, 28, 659-671.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1988). *Computer applications to curriculum-based measurement: Effects of teacher feedback systems*. Unpublished

- manuscript, Peabody College, Vanderbilt University, Nashville, TN.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education, 33*(1), 18-28.
- Ginsburg, H. P., & Allardice, B. S. (1984). Children's difficulties with school mathematics. In J. Lave & B. Rogoff (Eds.), *Everyday cognition: Its development in social context* (pp. 194-219). Cambridge, MA: Harvard University Press.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Needham Heights, MA: Allyn & Bacon.
- Good, R. H., & Jefferson, G. (1998). Contemporary perspective on curriculum-based measurement validity. In M.R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York: The Guilford Press.
- Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 25-49). Cambridge, MA: The MIT Press.
- Howell, K. W., Fox, S. L., & Morehead, M. D. (1993). *Curriculum-based evaluation: Teaching and decision making* (2<sup>nd</sup> ed.). Pacific Grove, CA: Brooks/Cole.
- Jordan, N. C. (1995). Clinical assessment of early mathematics disabilities: Adding up the research findings. *Learning Disabilities Research & Practice, 10*(1), 59-69.
- Kaminski, R. A., & Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 28*, 215-227.
- Kaminski, R. A., & Good, R. H. (1998). Assessing early literacy skills in a problem solving model: Dynamic Indicators of Basic Early Literacy Skills. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York: The Guilford Press.
- Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., Sanders, M., & Chiesa, J. (1998). *Investing in our children: What we know and don't know about the costs of and benefits of early childhood interventions*. Washington, DC: RAND.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: The Guilford Press.
- National Research Council. (1989). *Everybody counts*. Washington, DC: National Academy Press.
- Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development, 61*, 27-59.
- Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 Mathematics Report Card for the Nation and the States*. Washington, DC: National Center for Education Statistics.
- Resnick, L. B. (1989). Defining, assessing, and teaching number sense. In J. T. Sowder & B. P. Schappelle (Eds.), *Establishing foundations for research on number sense and related topics: Report of a conference* (pp. 35-39). San Diego: Center for Research in Mathematics and Science Education, San Diego State University.
- Riley, R. W. (1997). *Mathematics equals opportunity*. District of Columbia, U.S.: Federal Department of Education. (ERIC Document Reproduction Service No. ED 415119)
- Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7<sup>th</sup> ed.). Boston, MA: Houghton Mifflin Company.
- Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York: The Guilford Press.
- Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York: The Guilford Press.
- Skiba, R., Magnusson, D., Marston, D., & Erickson, K. (1986). *The assessment of mathematics performance in special education: Achievement tests, proficiency tests, or formative evaluation?* Minneapolis: Special Services, Minneapolis Public Schools.
- Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the Pine County norms: A compilation of findings* (Research Report No. 132). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.
- U.S. Department of Labor. (1990). *Work force 2000*. Washington, DC: Government Printing Office.
- U.S. Department of Labor, Bureau of Labor Statistics. (1997). *Occupational outlook handbook*. Washington, DC: U.S. Government Printing Office.
- VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*, 363-382.
- Woodcock, R. M., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Ben Clarke received his doctorate from the School Psychology Program at the University of Oregon in 2002. Currently he is a research associate at Pacific Institutes of Research. His primary research interests are in the areas of data-based decision making and early mathematics assessment and instruction.

Mark Shinn currently is a Professor of School Psychology at National-Louis University and a consultant to schools and state departments of education. From 1984-2003, he was Professor of School Psychology and Special Education at the University of Oregon. He is the recipient of the APA Division 16 2003 Jack Bardon Award for Distinguished Service and was recently appointed to the Review Panel for the Office of Special Education Programs Progress Monitoring Technical Assistance Center. His interests remain progress monitoring and use of CBM in a needs-based problem-solving service delivery system.