

Validation and Decision Accuracy of Early Numeracy Skill Indicators

Scott A. Methe
East Carolina University

John M. Hintze
University of Massachusetts at Amherst

Randy G. Floyd
The University of Memphis

Abstract. The purpose of this study was to develop short-duration assessment measures hypothesized to be valid samples of early mathematical behavior. The Early Numeracy Skill Indicators were designed using a curriculum-based assessment approach. Participants included 64 kindergarten children from a school district in the rural Northeast. Design components featured longitudinal correlation analyses conducted over a 26-week period. Decision analyses were completed using receiver operating characteristic techniques. Results indicated that selected Early Numeracy Skill Indicators tasks produced reliable, valid, and diagnostically accurate scores in relation to established criterion measures. Implications focus on the use and development of the measures as a means to prevent failure and enhance mathematics competency. Further, these tools complement the growing availability of early mathematics curriculum-based measures.

Preventing early learning problems through effective formative assessment has been shown to promote student success and enhance competence (Black & Wiliam, 1998; Daly, Hintze, & Hamler, 2000; Deno, 1989; Fuchs & Fuchs, 2001; Shinn, 1995). Screening young children for readiness skills upon school entry is an established educational tradition continually in need of empirical support (Fuchs & Fuchs, 2001; Gredler, 1992). Many schools feature kindergarten screening processes as means to establish relationships, examine entry-level skills, and prepare instruc-

tional support mechanisms. Typically, screening entails the brief examination of children for skill proficiency in foundational areas of academic development such as social behavior, speech, language, early literacy, and early numeracy (Howell & Nolet, 1999). The aim of these examinations is to help educators make accurate placement and instructional support decisions before the onset of schooling (Gredler, 1992).

Adopting a data-driven, curriculum-based measurement (CBM) assessment approach at kindergarten entry may reduce the

Correspondence regarding this article should be addressed to Scott A. Methe, East Carolina University, School Psychology Program, 104 Rawl Building, Greenville, NC, 27858; E-mail: methes@ecu.edu

Copyright 2008 by the National Association of School Psychologists, ISSN 0279-6015

need for higher stakes assessment later in a child's educational career and may serve to equalize opportunity (Black & Wiliam, 1998; Fuchs & Fuchs, 1986, 2001). Stanovich (1986) applied the idea of a "Matthew effect" to children learning to read; he noted that children with strong early literacy skills increasingly outperform less competent peers throughout their educational careers (Good, Simmons, & Smith, 1998). Early assessment and intervention are crucial for preventing this divisive trend. Large-scale data from the National Assessment of Educational Progress and the Early Childhood Longitudinal Study suggest similar Matthew effects. The National Assessment of Educational Progress is a nationally representative assessment examining subject matter proficiency. Fourth-grade results in mathematics from 1996 through 2004 indicate that the percentage of children at or below basic levels of proficiency (approximately 40%) has remained stable, whereas the number of children at or above proficient levels has doubled (National Center for Education Statistics, 2004). These data are consistent with the spirit of the Matthew effect in reading, suggesting that students with stronger skills are more likely to profit from education than lower skilled peers (Kavale, Forness, & Siperstein, 1999; Stanovich, 1986).

Factors placing young children at risk for mathematics achievement problems are evident in the results of the Early Childhood Longitudinal Study, which followed a cohort of 10,500 kindergarten students from 1998 to their third-grade year in 2002 (Denton & West, 2002). On a standardized measure of number sense, operations, and geometry designed to align with popular curricular "strands" (National Council of Teachers of Mathematics, 2000, 2006), students with no reported risk factors significantly outperformed students with more than one risk factor. Lower achieving children in third grade also began kindergarten with considerably fewer entry-level skills. This latter issue is consistent with the findings of Griffin, Case, and Siegler (1994), who demonstrated that experiential differences in number skills are less obvious upon kindergarten entry but be-

come more evident with the passage of time. Consistent with the intention of a response to intervention paradigm, these results are an alarming call to implement prevention efforts through early screening that informs the allocation of instructional resources.

Deno (1989) referred to fundamental math skills as *cultural imperatives*, a notion shared by numerous mathematics researchers and educators (Ginsburg & Russell, 1981; Griffin et al., 1994). Recognizing and using numbers is both a cultural practice and critical basic skill set (Baroody, 2004; Joram, Resnick, & Gabriele, 1995). As such, the National Council of Teachers of Mathematics (2006) has recently adopted a pared-down set of goals and objectives, termed *curriculum focal points*. At the kindergarten level, these standards place more emphasis on the number sense and operations strand, which serve as fundamental building blocks for developing more comprehensive skills over time (Clements, 2004; Clements, Sarama, & DiBiase, 2004). Further emphases are placed on building basic number skills, describing shapes, and assigning numbers to dynamic natural phenomena such as time, weather, and money.

Contrasting cultural imperatives with the epidemic of mathematics underachievement in American schools, problem-solving school psychologists are in great need of strategies that help to solve "big problems" through prevention and assessment-driven intervention (Shapiro, 2006). An instructionally relevant assessment approach should be informed by a sound theoretical and practical base with professional consensus on target content and domains. As a result of the Conference on Early Math Standards (Clements, 2004; Clements et al., 2004), early numeracy has such a consensual base, which is known as number sense. Similar to the role of phonological awareness in reading, number sense presupposes an awareness of the central conceptual structure of number (Griffin et al., 1994). Like a letter, a number is representative of a concept; instead of a language concept, a number represents a more objective percept. For example, the number 4 can represent four items, as well as two sets of two items. Given

this cognitive base, it is not surprising that number sense is defined by a relatively clear subset of numeric knowledge represented by a *mental number line* (Baroody, Ginsburg, & Waxman, 1983; Clements, 2004; Ginsburg & Russell, 1981; Griffin et al., 1994). When children are able to count, represent objects through one-to-one correspondence, understand relative size and polarity, and understand the ordinal nature of number, they are equipped with what Ginsburg and Russell (1981) term “informal” knowledge. Formal mathematical knowledge refers to the symbolic system—numbers—typically introduced upon entry to kindergarten. Children demonstrating difficulty in early mathematics have been shown to lack accuracy and fluency with these numerical concepts as well as with basic arithmetical operations (Gersten & Chard, 1999). The thrust of the research into mathematics CBM has targeted basic facts and operations (Foegen, Jiban, & Deno, 2007).

T1
AQ: 1-2

Outlined in Table 1, numerous early mathematics measures have begun to tap into target domains relevant to number sense (National Council of Teachers of Mathematics, 2000, 2006). In referencing virtually all available measures, tasks such as number recognition, counting, and enumeration of sets appear to exemplify important early math skills. Further, the results of the studies including these tasks suggest the importance of using a CBM approach to inform intervention for children with different levels of educational need. To facilitate assessment to intervention links, most of the available measures in Table 1 include stimulus features that are instructionally relevant. For example, at basic levels of knowledge, children are directed to name, identify, produce, select, and choose (Kame’enui & Simmons, 1990). These measures also vary in terms of the use of numerals versus pictures or sets of objects.

In the current study, we intend to augment the content of the current research base in three ways. First, this study includes replications of selected assessment tasks (e.g., number recognition). Second, we include variations of existing tasks (e.g., matching numbers to sets of objects). Third, the current

study proposes a set of unique assessment tasks linked to both informal and formal early mathematical curricular domains such as counting up, ordinality, and numeral recognition (Aubrey, 2001; Clements, 2004; National Council of Teachers of Mathematics, 2006). These domains are featured in Clements (2004) as hierarchical concepts resembling key “big ideas” for early mathematics: counting, ordering, partitioning, decomposing, grouping, and basic operations (see Methe & Riley-Tillman, in press). The Early Numeracy Skill Indicator (ENSI) tasks were designed to operationalize these ideas.

AQ: 3

In addition, this study extends the methodological content of the current research base by including receiver operating characteristic (ROC) curves. Consistent with previous research (Hintze, Ryan, & Stoner, 2003; Hintze & Silbergitt, 2005; Silbergitt & Hintze, 2005), ROC analyses were used in this study to examine (a) the accuracy of screening decisions made on the basis of cutoff scores (Streiner, 2003; Swets, 1992), (b) the efficiency in generating indices for a wide range of cutoff scores, and (c) the ability to examine the utility of the ranges of scores in the form of visual plots. To complete ROC analyses, it is necessary to obtain scores on an experimental screening measure concurrently with binary-coded criterion scores delineating the presence or absence of a condition. In this study, a math problem is represented by an obtained standard score at or below the 24th percentile. Using the 25th percentile to identify low achievers is a recommended and commonly accepted cutoff that has been used in kindergarten screening as well as extensive longitudinal studies in reading diagnostics (Gredler, 1992; Shaywitz, Fletcher, Holahan, & Shaywitz, 1992). With reference to each obtained score on an experimental measure, ROC analyses code into four indices: (a) true positives (1 and 1), (b) true negatives (0 and 0), (c) false positives (1 and 0), and (d) false negatives (0 and 1). Each experimental score is automatically assigned a set of statistics, informing the user of its decision accuracy.

The central purpose of the current research is to assist the developing paradigm of

Table 1
Review of Early Mathematics CBM Measures

| Study | Grade and Sample | Activities Involved in Tasks | Technical Summary |
|--|------------------|---|--|
| Joyce & Wolking (1987) | K-1 (144) | Count dots Name printed numbers Count backwards from ten | Did not use criterion to validate. No reliability data. Classification accuracy ranged from 76% to 78% correct classification. |
| Daly, Wright, Kelly, & Martens (1997) | 1 (30) | Name printed numbers Count orally from one Count orally from one Write dictated numbers Select dictated number Name shapes | Reliability analyses indicate utility for screening. Moderate concurrent and predictive correlations between CBA probes and criteria. |
| Reid, Morgan, DiPerna, & Lei (2006) | PK (117) | Count orally from 1 Name printed numbers Count objects aloud Identify number of objects in set Identify patterns using shapes Identify relative size | Strong item and scale reliability analyses, utility for progress decisions. Moderate to strong concurrent correlations. |
| Floyd, Hojnoski, & Key (2006) | PK (163) | Count objects aloud Count orally from 1 Name printed numbers Compare and name larger quantity | Strong test-retest reliabilities. Moderate to strong concurrent correlation. Factor analyses indicate single-factor loading. |
| VanDerHeyden, Broussard, & Cooley (2006) | PK, K (102, 82) | Choose correct dictated number Count objects aloud Count aloud Discriminate various symbols Count circles then identify number | Moderate to strong concurrent correlations across probes. Measures detected predicted performance differences. |

Note. CBM = curriculum-based measurement; CBA = curriculum-based assessment. See Foegen, Jiban, and Deno (2007) for a more comprehensive technical review.

early mathematics assessment by proposing experimental measures linked to important domains of early mathematical knowledge (National Council of Teachers of Mathematics, 2000, 2006). Given our interest in the domain of early mathematics, the research methods are designed to directly examine the reliability and to produce validity evidence based on external relations for a variety of experimental measures targeting early numeracy. For selected tasks of the ENSIs, our experimental measures, we hypothesize the following: (a)

stability in obtained scores across time, examiners, and within items; (b) meaningful relationships between experimental and criterion measures; and (c) experimental measures will accurately predict risk status on criterion measures when used in reference to cutoff scores.

Method

Participants and Setting

Children included in the full longitudinal analyses were 64 kindergarten students (30

boys and 34 girls) enrolled in a public school district in the rural northeastern United States. The school district included 10 kindergarten classrooms in three separate elementary schools, with a total kindergarten enrollment of 139 students. Data were collected during the 2003–2004 school year. Each measure listed in the Materials and Measures section (experimental measures, criterion measures, and teacher rankings) had been adopted by the participating school district as part of its kindergarten screening protocol; therefore, no formalized consent was necessary. As such, the current study was exempt from review according to the procedures of the university's human subjects institutional review board. Initial study participants ($N = 100$) were selected using a random number generator. At the beginning of the study, participants' ages ranged from 4 years 10 months to 6 years 8 months, with a median age of 5 years 1 month.

AQ: 4

AQ: 5

From these initial 100 students, 13 were selected to participate in pilot testing of the experimental measures, 8 students moved out of the district, and full data on the remaining 15 students were incomplete owing to other factors such as student absences and lack of personnel. Demographic characteristics of the 64 children included in the full analysis were similar to those of the full kindergarten ($N = 139$) as well as kindergarten through twelfth grade district data (47% boys; 53% girls; 93% Caucasian, 4% Hispanic, 3% African American; 49% free and reduced-priced lunch). Comparisons of the current sample and district demographics to state free and reduced-priced lunch rates (27%) indicated higher rates. However, sample and district data compared somewhat more evenly to 2004 U.S. census data, indicating that close to 35% of school-aged children meet poverty threshold criteria when such criteria are doubled to account for aggregate family size.

Materials and Measures

Test of Early Mathematics Achievement, Third Edition (TEMA-3). The mathematics ability of the kindergarten sample was assessed using the TEMA-3 (Ginsburg & Ba-

roody, 2003). The TEMA-3 was chosen because of its sound theoretical base; the authors of the TEMA-3 have been widely involved in the development of instruction, assessment, and curriculum at both practice and policy levels (Baroody, 2004; Baroody et al., 1983; Clements, 2004; Clements et al., 2004; Ginsburg & Baroody, 2003; Ginsburg & Russell, 1981). The TEMA-3 measures numbering skills, number-comparison facility, numeral writing and recognition, number facts, calculation skills, and number concepts. These skills comprise the "number sense and operations" curricular strand (National Council of Teachers of Mathematics, 2000). The TEMA-3 is available in two alternate forms. Reported internal consistencies were above .92, and the alternate-form and two-week test-retest reliability both exceeded .80. Concurrent correlation coefficients between the TEMA-3 and four commonly used published norm referenced tests ranged from .55 to .91. Standard scores were used as the primary criterion measure due to existing evidence of validity, to control on age effects, and to represent best the "true" population variance (Allen & Yen, 1979).

Teacher ratings. Teachers rated the math performance of each child following the fall and spring test administration. Teacher ratings were included as a relatively efficient additional criterion measure as well as a means of social validation (Gresham & MacMillan, 1997). In a comprehensive review of the literature pertaining to teacher judgments of achievement, Perry and Meisels (1996) concluded that teachers are highly accurate in predicting which students are actually at risk, using common achievement tests as criteria. Teachers were asked to rate student performance at three levels of mastery in the kindergarten curriculum (0 = mastered none or few curriculum objectives; 1 = mastered some of the curricular objectives, and 2 = mastered all or most of the curricular objectives).

Experimental measures. Based on an extensive literature review, a set of four measures were constructed to operationalize kin-

T2

dergarten mathematics knowledge, often referred to as “number sense” (Chard et al., 2005; Clarke & Shinn, 2004; Clements, 2004; Clements et al., 2004). Collectively, these measures are referred to as the Early Numeracy Skill Indicators (ENSIs). Typically, number sense includes free counting, enumerating sets of items, identifying attributes of objects, naming numbers, and using order in counting numbers. In constructing the ENSIs, both non-numerical and numerical knowledge were included. Baroody (Baroody et al., 1983; Baroody, 2004), and Ginsburg and Baroody (2003) classify this type of knowledge as formal and informal, suggesting that they interface upon kindergarten entry. Descriptions of the four experimental measures are presented in Table 2, and include Counting-on Fluency (COF), Ordinal Position Fluency (OPF), Number Recognition Fluency (NRF), and Match Quantity Fluency (MQF) measures. Each measure was administered individually and under timed conditions to assess both response accuracy and fluency. Each measure included a set of standardized instructions to account for student hesitations, misunderstandings, and mistakes. Administration directions also included practice items with a prompt to begin each probe. The OFF task included both production- and selection-type responses, whereas the other tasks required either production or selection tasks. Each obtained score is presented as a rate-based metric, representing units correct per minute.

Procedure

Procedural integrity. To facilitate comparisons among scores and proactively account for administration error, three methods of procedural integrity were included in the current study: (a) examiner training, (b) observations, and (c) pilot test administration. Graduate students with experience in administering CBM to children served as primary examiners. Before each testing period, examiners were provided with practice and training in the procedures of the testing, supervision, observation, and feedback. Comprehensive 3-hr sessions were provided by the primary author 1

day before the fall and spring assessment periods. Session attendees were the primary author and seven graduate students. During the session, the primary author presented both the TEMA-3 and ENSI tasks to trainees, and modeled the administration of the tests. The remainder of the session was dedicated to administering the test to a partner with feedback from the primary author. Session goals were to familiarize participants with the materials and procedures. Brief 1-hr sessions with observational feedback were provided before the winter sessions. Before each administration period, examiners took a 10-item quiz regarding test administration procedures, directions, and discontinue rules. All examiners scored 100% on the quiz before test administration.

Observation of testing procedures.

The district pupil services coordinator, who typically administers the kindergarten screening protocols, observed the fall testing session and completed two assessment integrity observations per each test administrator. Behaviors to be observed included (a) stating directions verbatim, (b) speaking clearly, (c) using the stopwatch as intended, and (d) adequate pacing of tasks. Despite directions for numeric coding on an assessment integrity rating scale, the coordinator opted for brief qualitative comments. Examination of notes recorded by the coordinator indicated that each examiner accurately stated directions, used the stopwatch as intended, spoke clearly, and proceeded at an adequate pace.

Pilot administration. To field test the experimental measures, 13 students taken from the original sample of 100 students were administered each ENSI task. Consequently, these 13 students were not included in the final sample. The purpose of the pilot testing was to examine the measures for consistency and clarity in directions, ease of use, and ease of scoring. Three of the trained graduate students including the current author administered the tests to the 13 students. Following the administration, a group feedback session was held and revisions recommended in (a) directions and testing prompts and (b) scoring sheet lay-

Table 2
Description of ENSI Measures

| Measure | Stimulus Materials | Directions | Early Mathematics Domain Assessed |
|---------|---|---|---|
| COF | None | “I want you to count some numbers aloud. When I say begin, start at the number I tell you, and count to another number I tell you. Listen carefully as I try one. I will start at four and count to seven. I say ‘four, five, six, seven.’ Your turn. Start with three and count to six.” | COF assesses an “unbroken chain” of the counting sequence (Clements, 2004) |
| MQF | Card with a group of items on the left and a set of numerals on the right. | “We are going to look at some pictures and numbers. I want you to point to the number over here that tells how many things are over here. Let’s try one. Watch me. I am going to point to the number 4. It looks like there are four ice cream cones in this circle. The ice cream cones here are the same as the number 4, so I point to the number 4. Ready, your turn! Let’s try one with cats. I want you to point to the number that tells how many cats are in the circle.” | Enumeration; formal written representation of sets (Clements, 2004) |
| NRF | Card with set of randomly selected numbers ranging from 1 to 20. | “I want you to say the names of these numbers as fast as you can. I want you to start here, and name all the numbers in this row before you go on to the next row. Name as many numbers as you can. Just do your best and tell me the names of the numbers. Put your finger on the first number. Ready, begin.” | Ability to identify numerals to 20 (NCTM, 2000) |
| OPF | Card with five objects in a straight line (cat, cookie, toaster, and so on) and dots under each item. | “We are going to look at some pictures. Watch me. I will name the pictures and say what place they are in. The _____ is in first place, the _____ is in second place . . . I want you to tell me what place the ball is in. Good. Now, point to the picture in 4th place.” | Order, structure, and dynamism of numbers (Clements, 2004; Griffin et al., 1994). |

Note. ENSI = Early Numeracy Skill Indicators; COF = Counting-on Fluency; MQF = Match Quantity Fluency; NRF = Number Recognition Fluency; OPF = Ordinal Position Fluency; NCTM = National Council of Teachers of Mathematics.

out. Minor revisions for language clarity and consistency in the use of prompts were made.

Data collection. Data on the experimental measures were collected in the fall, winter, and spring. The TEMA-3 and teacher rating criterion data were collected in the fall and spring to facilitate concurrent and predic-

tive validity analyses. Administration of all measures occurred outside each kindergarten classroom, where each student sat at a desk facing the administrator in relatively close proximity to other children in the same conditions. Teachers were asked by building principals to minimize transitions and disruptions during testing. Data were collected individu-

Table 3
Descriptive Statistics for ENSI and TEMA-3 Measures Across Fall, Winter, and Spring Assessment Periods

| Measure | Fall | | Winter | | Spring | |
|---------|----------|-----------|----------|-----------|----------|-----------|
| | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> |
| COF | 7.8 | 4.1 | 10.4 | 3.6 | 12.8 | 3.7 |
| NRF | 24.0 | 16.3 | 32.8 | 16.6 | 42.1 | 16.9 |
| MQF | 17.8 | 7.4 | 19.8 | 6.6 | 21.1 | 6.3 |
| OPF | 22.8 | 21.9 | 33.0 | 19.9 | 38.6 | 19.2 |
| TEMA-3 | 93.4 | 13.0 | — | — | 102.8 | 12.4 |

Note. ENSI = Early Numeracy Skill Indicators; TEMA-3 = Test of Early Mathematics Achievement, Third Edition; COF = Counting-on Fluency; MQF = Match Quantity Fluency; NRF = Number Recognition Fluency; OPF = Ordinal Position Fluency. ENSI scores presented in raw-score units. TEMA-3 scores are presented as standard scores ($M = 100$, $SD = 15$).

ally from each child directly outside the student's classroom. The sessions were designed to take 2 weeks, with administration of the ENSI measures first, followed by the TEMA-3 administration. The first data collection period occurred in mid-October with subsequent test sessions equally spaced 13 weeks apart. Examiners scored each test immediately following its administration. The order of ENSI task administration was randomized during the winter and spring data collection periods to dismiss order effects. Using a random number generator, the numbers 1–4 were matched to each ENSI task, uniquely varied in sets, and each assessment packet was assembled with respect to these numbers. To ensure random assignment, each examiner was first given a unique number from 1 to 7; a random number generator was used to produce a single number (among 1 to 7) while assembling each packet. The examiner's name was placed on the assessment packet before the test. On the day of testing, the examiner retrieved the premade packets and proceeded to assess the specified child.

Results

Descriptive Statistics and Data Screening

Before analysis, we examined scoring accuracy, the accuracy of the data file, and the degree to which obtained data were normally

distributed (Tabachnick & Fidell, 2007). Scoring accuracy was calculated by dividing the number of corrections by the total number of tasks administered. Scoring accuracy was high for both the ENSIs (>95%) and TEMA-3 (>89%). Accuracy of the data file was checked with two proofreaders, who checked the original data from the test protocols against the data entered into the file. Means, standard deviations, and skewness and kurtosis statistics were examined for plausibility. Notably high standard deviations were evident in the fall OPF and NRF measures. Although kurtosis was within acceptable limits across ENSI measures, slight negative skew was apparent in selected tasks. Scores on the experimental and criterion measures were converted to a z distribution, and a criterion of 3.29 standard deviations (Tabachnick & Fidell, 2007) was the cutoff for outliers. Six outlying scores were detected over the course of the study across ENSI measures. These scores were deleted and mean scores were substituted. Inserting mean scores resulted in skew reduction but did not affect correlation coefficients. (See Table 3.)

Reliability

Results of internal consistency and test-retest reliability analysis are included in Table

AQ: 6

AQ: 7
T3

Table 4
Internal Consistency and Test–Retest Reliability Coefficients for ENSI Subtests

| Subtest | Internal Consistency | Test–Retest |
|---------|-------------------------|---------------------------|
| | Mean KR-20 ^a | 13 Weeks (<i>n</i> = 20) |
| COF | .80 (78) | .68 |
| NRF | — | .98 |
| MQF | .53 (81) | .74 |
| OPF | .83 (77) | .81 |

Note. ENSI = Early Numeracy Skill Indicators; COF = Counting-on Fluency; MQF = Match Quantity Fluency; Sample size in parentheses. NRF = Number Recognition Fluency; OPF = Ordinal Position Fluency; KR-20 = Kuder-Richardson formula 20. Sample size in parentheses. ^aRanges for KR-20 coefficients = COF, .75–.85; MQF, .40–.75; OPF, .80–.85.

T4

4. Internal consistency estimates are reported as α (20), utilizing the Kuder-Richardson procedure (KR 20; Allen & Yen, 1979). Alpha coefficients such as the KR 20 are not recommended if the speed of a test prevents students from attempting every task, thus differentially affecting any given dichotomous answer (Allen & Yen, 1979). With the exception of the NRF task, latency-type administration procedures allowed students to attempt every item, and item-level performance resulted in a binary (i.e., correct or incorrect) score. For the internal consistency reliability analysis, forms were excluded in which students received a score of 0 correct from the discontinue rules, which prohibit task completion. Using Pearson product-moment correlation, stability coefficients—representing test–retest reliability estimates—were obtained from data collected over 13 weeks (from the winter to spring session). Data were analyzed on the same examiner–student pairs across sessions to minimize variation because of the examiner (*n* = 20).

Internal consistency reliability estimates were high for the COF and OPF tasks (.80 and .83, respectively), but it was low for MQF (.53). Stability coefficients were high for two

ENSI measures (.81 for OPF and .98 for NRF). The MQF measure evidenced lower stability, with a coefficient just under .80, and the least reliable measure was COF (.68). Using internal consistency and stability coefficients as converging evidence, reliabilities were high for the NRF and OPF measures—with lower reliability over time evidenced on the COF measure and the lowest reliability for the MQF measure. To examine consistency across persons, a one-way analysis of variance was conducted using examiner as the grouping variable. On the fall and spring administration of the TEMA-3, no significant differences across examiners were detected ($F = 0.324, p < .05$). Differences were detected among examiners only on the fall OPF measure ($F = 5.884, p < .05$) and the spring NRF measure ($F = 2.48, p < .05$).

Validity

Examining the Pearson bivariate correlations in Table 5, the NRF, OPF, and COF evidenced moderate concurrent relations with the TEMA-3 scores during the fall and spring assessment periods. All coefficients were .50 or higher. Although MQF also demonstrated moderate relations with the TEMA-3 score in the fall, its concurrent correlations were weak ($r = .20$) in the spring, indicating a coefficient of determination at .04 (Allen & Yen, 1979). Relations between teacher ratings (TR in Table 5) and ENSI measures were examined using the eta (η) statistic for cross-tabulated nominal by interval scales. Correlations were moderate to strong across ENSI measures and assessment periods (range = .66–.89). Teacher rating data were also strongly related to the TEMA-3 in the fall and spring (.83 and .86, respectively).

Additional correlation coefficients were calculated to examine predictive relations between ENSI measures and the TEMA-3 and teacher ratings collecting in the spring. Results are included in Table 6. The NRF measures from the fall and winter administrations demonstrated the strongest predictive correlations with the TEMA-3. The OPF, COF, and MQF measures also evidenced moderate predictive

T5

T6

Table 5
Correlations Between ENSI Measures, TEMA-3 Measures,
and Teacher Ratings

| Measure | 1 | 2 | 3 | 4 | 5 | 6 |
|-------------------------|-----|-----|-----|-----|-----|---|
| Fall (<i>N</i> = 77) | | | | | | |
| 1. COF | — | | | | | |
| 2. NRF | .47 | — | | | | |
| 3. MQF | .36 | .49 | — | | | |
| 4. OPF | .42 | .71 | .45 | — | | |
| 5. TEMA-3 | .50 | .72 | .55 | .63 | — | |
| 6. TR ^a | .68 | .89 | .70 | .81 | .86 | — |
| Spring (<i>N</i> = 64) | | | | | | |
| 1. COF | — | | | | | |
| 2. NRF | .56 | — | | | | |
| 3. MQF | .35 | .37 | — | | | |
| 4. OPF | .49 | .64 | .38 | — | | |
| 5. TEMA-3 | .55 | .64 | .20 | .60 | — | |
| 6. TR ^a | .70 | .89 | .66 | .79 | .83 | — |

Note. ENSI = Early Numeracy Skill Indicators; TEMA-3 = Test of Early Mathematics Achievement, Third Edition; COF = Counting-on Fluency; NRF = Number Recognition Fluency; MQF = Match Quantity Fluency; OPF = Ordinal Position Fluency; TR = teacher ratings.

^aTR uses a cross-tabulated eta (η) statistic for nominal by interval scale.

correlations with the TEMA-3. Using teacher ratings as the predictive outcome, fall performances on the NRF (.87), OPF (.79), and MQF (.72) demonstrated strong relations with spring teacher ratings of curricular mastery, whereas winter performances on NRF (.88), OPF (.77), and COF (.70) were strong predictors.

Diagnostic Accuracy

Diagnostic accuracy results are displayed in Table 7. Six indices are used to describe the nature of the 2×2 decision matrix: (a) sensitivity, (b) specificity, (c) positive prediction, (d) negative prediction, (e) correct classification, and (f) kappa. Among all students with a math problem (e.g., at or below the 24th percentile on the TEMA-3), a cutoff score is sensitive as the percentage of true positive classifications approaches 100%; given a negative test, a score has specificity as true negative classifications also approach 100%. Positive and negative predictive indices are the proportion of cases falling above and

below the specified cut score, respectively, that subsequently meet criteria for classification on the TEMA-3. Kappa measures the strength of association between selected measures and the TEMA-3 with respect to the proportion of chance agreement that exists in any classification decision (Watkins & Pacheco, 2000).

Cutoff scores on the NRF and OPF measures suggested adequate levels of sensitivity and specificity consistent with recommended levels in the range of .75 (Swets, 1992). Evident across cutoff scores were higher indices of negative prediction. This indicates ENSI cutoff scores are much stronger when used to rule out risk status in students who actually do not have the problem. However, using the OPF cutoff score of 15, specificity indices remind the user that there remains a .30 probability that any student nominated as low risk may be falsely classified. Remediating this problem calls for increasing the cutoff score, resulting in a trade-off with the number of false positives. That is, not all of the cases

Table 6
Predictive Validity of ENSI Subtests
With Spring Criterion Measures
(N = 64)

| Measure | Spring TEMA-3 | Spring Teacher Ratings ^a |
|---------|------------------|--|
| Fall | | |
| COF | .46 | .57 |
| NRF | .70 | .87 |
| MQF | .41 | .72 |
| OPF | .58 | .79 |
| Winter | | |
| COF | .62 | .70 |
| NRF | .66 | .88 |
| MQF | .47 | .61 |
| OPF | .57 | .77 |

Note. ENSI = Early Numeracy Skill Indicator; TEMA-3 = Test of Early Mathematics Achievement, Third Edition; COF = Counting-on Fluency; NRF = Number Recognition Fluency; MQF = Match Quantity Fluency; OPF = Ordinal Position Fluency.

^aCross-tabulated eta (η) statistic for nominal by interval scale.

below the raised cutoff criterion will actually be at risk. The cutoff scores selected for display in Table 7 were those demonstrating an optimal balance between sensitivity and specificity.

Ranked from high to low, NRF, OPF, COF, and MQF produced indices ranging from 84% to 58% correct classification. To adjust for chance agreement in hit rates and to further examine general levels of agreement, kappa indices can be interpreted much like correlation coefficients (Watkins & Pacheco, 2000). Good to moderate relations characterizing overall decision accuracy were notable in the OPF and NRF measures. Moderate to poor kappa coefficients were notable for the MQF and COF subtests, indicating that the range of cutoff scores, in general, failed to agree with the TEMA-3 as it delineated risk status.

Developmental Sensitivity

The utility of a CBM also hinges upon its ability to model change over time. It is thus

important for a measure to model change quantifiably (Clarke & Shinn, 2004; Fuchs, Fuchs, Hamlett, & Walz, 1993; Shinn, 1995). Descriptive statistics indicate that each measure indexed student growth over the course of a school year. The NRF and OPF measures allowed for quantification in whole raw-score units (approaching one unit per week). For example, over a 13-week period, NRF showed increments of .84 raw-score units per week. Conversely, COF indices changed by approximately .20 raw-score units per week. Consequently, growth increments for COF are small and raw-score units, although fundamental to CBM, may be inappropriate for indexing change using this task (Fuchs et al., 1993).

Discussion

With regard to each of the research hypotheses, the OPF and NRF measures evidenced reliability, validity, and produced cut-off scores that facilitated accurate classification decisions. Although each of the measures demonstrated relatively accurate grouping decisions, the remaining two ENSI measures (COF and MQF) evidenced lower levels of reliability and validity. This study is aligned with the recommendation of Fuchs (2004), who suggested examining technical properties of the static score using methods such as criterion validation and reliability analyses. Our results yield three key highlights of the OPF measure: (a) adequate technical properties, (b) little overlap with previously tested CBM probes, and (c) relevant content that examines knowledge of ordinal numbers, which help children understand the structured yet dynamic nature of number and the number line (Baroody, 2004; Clements, 2004; National Council of Teachers of Mathematics, 2006). Regarding the second point, the degree to which the OPF task accounts for unique variance in early mathematics achievement is an empirical question open for further inquiry and development.

One limitation to this study was our failure to recommend an optimal combination or sequence of measures, primarily because of the boundaries of our initial hypotheses. Al-

Table 7
Diagnostic Accuracy Statistics for Selected Cutoff Scores

| Measure | Cutoff Score | Sensitivity | Specificity | PPP | NPP | CC | κ |
|---------|--------------|-------------|-------------|-----|-----|-----|----------|
| OPF | 14 | .75 | .75 | .64 | .83 | .75 | .48 |
| | 15 | .83 | .70 | .63 | .88 | .75 | .50 |
| NRF | 12 | .71 | .93 | .85 | .84 | .84 | .66 |
| | 19 | .79 | .78 | .68 | .86 | .78 | .55 |
| COF | 7 | .71 | .73 | .61 | .81 | .72 | .42 |
| | 8 | .75 | .48 | .46 | .76 | .58 | .20 |
| MQF | 16 | .75 | .68 | .58 | .82 | .70 | .40 |
| | 17 | .79 | .65 | .58 | .84 | .70 | .41 |

Note. PPP = positive predictive power; NPP = negative predictive power; CC = percentage of cases correctly classified; OPF = Ordinal Position Fluency; NRF = Number Recognition Fluency; COF = Counting-on Fluency; MQF = Match Quantity Fluency.

though we cited theoretical data indicating differences between prenumeric and numeric knowledge, our analyses yielded to the lack of applied data pertaining to (a) the exact sub-skills that best exemplify number sense and (b) the recommended sequence of such skills. Developing measures that model growth in the short-term can help practitioners identify and treat skill breakdowns (Howell & Nolet, 1999). We believe these to be critical directions for future research into early numeracy CBM. To inform such directions, we direct readers to the extensive work included in Clements et al. (2004).

Given that the ENSI tests resemble sub-skill mastery measures, which are assessment tasks with specific short-term objectives (Fuchs & Deno, 1991; Hintze, Christ, & Methe, 2006), it is expected that the normality of the experimental distributions would vary, represented by problems in skewness. Although this may prove a limitation to the results, inspecting non-normal histograms may be a helpful direction for research examining the sequencing of mathematics knowledge. For example, group progression on the OPF task suggests a measurable developmental trend. Visual analyses of these histograms indicate strong positive skew in the fall of kindergarten, normality in the winter, and strong negative skew in the spring. This type of data leads to inquiries about testing readiness and

how ENSI measures model sequences of early mathematics knowledge.

Clements' (2004) hierarchy of number sense and operation skills is a useful starting point for future research into diagnostic assessment using selected early numeracy CBM measures (Methe & Riley-Tillman, in press). Because this hierarchy operationalizes key domains relevant to number sense and operations strands included in many state curricula (National Council of Teachers of Mathematics, 2006), it can be thought of as a unique organizing sequence, similar to the means by which the Dynamic Indicators of Basic Early Literacy Skills (Good, et al., 1998) model the process of acquiring word reading facility. As such, research should endeavor to examine the fit of ENSI and other commercially available early numeracy measures into a skill progression key to the purposes of diagnostic assessment for pinpointing and remediating specific skill deficits. Because the MQF and COF measures in the current study failed to demonstrate strong technical features related to the static score, these and tasks from other studies should be better developed before inclusion in diagnostic research.

Another limitation to the current study was the 13-week intervals used to establish consistency over time. These intervals were lengthier than what is necessary to establish test-retest reliability. Linn and Gronlund

AQ: 8

(2000) indicate that test–retest intervals should be linked to the purposes of the test under development. Given that our purposes were to develop screening measures, we believe the 13-week interval to be acceptable for providing estimates of stability, however limiting it proves as a measure of reliability. Recommended methods to determine test–retest reliability rely on shorter time intervals in the absence of instruction (perhaps over a school vacation), and thus address variability related to factors extraneous to the test.

Although there were limitations to our use of the TEMA-3 as gold standard, it is important to note that no perfect criterion is available for any diagnostic condition (Swets, 1992). Instead, it is important to select a comprehensive sample of the phenomenon of interest. We included teacher ratings to address the threat imposed by a single criterion measure. Although it was not our intent to develop a scale for such ratings, we believe the consistency of our findings would have been enhanced if we endeavored to concurrently develop and validate a scale with clearer boundaries on curricular objectives. Such a rating scale could prove quite useful if carefully aligned with national, state, local and research-based standards.

Further limitations included (a) failure to counterbalance administration of the criterion and experimental measures, (b) lack of inter-rater reliability data, (c) evidence of some interexaminer inconsistency, and (d) attrition of participants. In regard to counterbalancing, given that the TEMA-3 was administered to the kindergarten group before administration of the ENSIs, we cannot rule out the alternative hypothesis that positive performances on the ENSIs were from practice effects. Further, we sought to maximize procedural integrity throughout the process of the study, but a second alternative hypothesis supposes that observed early math problems were actually from inconsistencies in the administration procedures. Although these results should be interpreted both provisionally and in tandem with reliability results, the data indicate a need for refined test administration procedures. Refined procedures and logistics are

also important in light of attrition in the current study.

In virtually any educational setting, risk status changes when cutoff scores are adjusted. One key highlight of CBM that has led to its viability in school assessment is the utility of its raw scores to guide the decision-making process using local norms. This idea has implications regarding the development of early numeracy assessment as a flexible paradigm useful for instructional provision. To illustrate, 100% of students scoring above a 23 on the OPF measure concurrently scored above the 25th percentile on the TEMA (perfect specificity). It is defensible to describe such performance as a means to define a low-risk group (ostensibly hinging on the idea that above the 25th percentile is synonymous with low risk—certainly not a consistent finding). Similarly, 100% of students who obtained an 8 concurrently scored below the 10th percentile on the TEMA (perfect sensitivity), thus defining a high-risk group. Scores falling between these parameters thus may indicate the need for supplementary support. Although this depiction cannot perfectly account for false classification and needs to consider base rates, the idea and methodology complement not only the content of an assessment battery, but the process by which school psychologists arrive at important educational decisions.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.
- Aubrey, C. (2001). Early mathematics. In T. David (Ed.), *Promoting evidence-based practice in early childhood education: Research and its implications* (pp. 185–210). Amsterdam: Elsevier Press.
- Baroody, A. J. (2004). The developmental bases for early childhood operations and number standards. In D. H. Clements & J. Sarama (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education*. Mahwah, NJ: Erlbaum.
- Baroody, A. J., Ginsburg, H. P., & Waxman, B. (1983). Children's use of mathematical structure. *Journal for Research in Mathematics Education*, *14*, 156–168.
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139–144.
- Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Prelim-

- inary findings. *Assessment for Effective Intervention*, 30(2), 3–14.
- 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.
- Clements, D. H. (2004). Major themes and recommendations. In D. H. Clements, J. Sarama, & M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education*. Mahwah, NJ: Erlbaum.
- Clements, D. H., Sarama, J., & DiBiase, M. (Eds.). (2004). *Engaging young children in mathematics: Standards for early childhood mathematics education*. Mahwah, NJ: Erlbaum.
- Daly, E. J., Hintze, J. M., & Hamler, K. R. (2000). Improving practice by taking steps toward technological improvements in academic intervention in the new millennium. *Psychology in the Schools*, 37, 61–72.
- Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1–17). New York: Guilford Press.
- Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade* (No. NCES 2002–125). Washington, DC: National Center for Education Statistics.
- FOEQ: 9
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education*, 41, 121–139.
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, 33, 188–192.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, 57, 488–500.
- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199–208.
- Fuchs, L. S., & Fuchs, D. (2001). Principles for the prevention and intervention of mathematics difficulties. *Learning Disabilities Research & Practice*, 16, 85–95.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Walz, L. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, 22, 27–48.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematical disabilities. *The Journal of Special Education*, 33, 18–28.
- Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability, Third Edition*. Austin, TX: Pro-Ed.
- Ginsburg, H. P., & Russell, R. L. (1981). Social class and racial influences on early mathematical thinking. *Monographs of the Society for Research in Child Development*, 46(6), 69–69.
- Good, R. H., III, Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *Educational and Child Psychology*, 15, 56–70.
- Greder, G. R. (1992). *School readiness: Assessment and educational issues*. Brandon, VT: Clinical Psychology Publishing.
- Gresham, F. M., & MacMillan, D. L. (1997). Teachers as 'tests': Differential validity of teacher judgments in identifying students at-risk. *School Psychology Review*, 26, 47–60.
- Griffin, S. A., Case, R., & Siegler, R. S. (1994). Right-start: 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: MIT Press.
- Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools*, 43, 45–56.
- Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review*, 32, 541–555.
- Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, 34, 372–386.
- Howell, K. W., & Nolet, V. (1999). *Curriculum-based evaluation: Teaching and decision making*. Belmont, CA: Wadsworth Publishing.
- Joram, E., Resnick, L. B., & Gabriele, A. J. (1995). Numeracy as cultural practice: An examination of numbers in magazines for children, teenagers, and adults. *Journal for Research in Mathematics Education*, 26, 346–361.
- Kame'enui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: The prevention of academic learning problems*. Columbus, OH: Merrill.
- Kavale, K. A., Forness, S. R., & Siperstein, G. N. (1999). *Efficacy of special education and related services*. Washington, DC: American Association on Mental Retardation.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall.
- Methe, S. A., & Riley-Tillman, T. C. (in press). An informed approach to selecting and designing early mathematics interventions. *School Psychology Forum*.
- National Center for Education Statistics. (2004). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (No. NCES 2005–464). Washington, DC: U.S. Government Printing Office.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through 8th grade mathematics: A quest for coherence*. Reston, VA: Author.
- Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?* (No. NCES Working Paper No. 96–08). Washington, DC: National Center for Education Statistics.
- Russell, R. L., & Ginsburg, H. P. (1984). Cognitive analysis of children's mathematics difficulties. *Cognition and Instruction*, 1, 217–244.
- AQ: 11
- AQ: 12

- Shapiro, E. S. (2006). Are we solving the big problems? *School Psychology Review, 35*, 260–265.
- Shaywitz, B. A., Fletcher, J. M., Holahan, J. M., & Shaywitz, S. E. (1992). Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut Longitudinal Study. *Journal of Learning Disabilities, 25*, 639–648.
- Shinn, M. R. (1995). Best practices in curriculum-based measurement and its use in a problem solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (3rd ed., pp. 547–567). Washington, DC: National Association of School Psychologists.
- Silbergliit, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*, 304–325.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–406.
- Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*, 209–219.
- Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon/Pearson Education.
- Wang, M. C., Resnick, L. B., & Boozer, R. F. (1971). The sequence of development of some early mathematics behaviors. *Child Development, 42*, 1767–1778.
- Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10*, 205–212.

AQ: 13

Date Received: November 15, 2006
 Date Accepted: May 27, 2008
 Action Editor: Sandra Chafouleas ■

Scott A. Methe is Assistant Professor of School Psychology at East Carolina University in Greenville, North Carolina. In 2005, he received his PhD in School Psychology from the University of Massachusetts at Amherst. His research interests include assessment and intervention in early mathematics, facilitating recreational reading habits in elementary-aged children, and the assessment literacy of educational professionals.

John M. Hintze is Associate Professor of School Psychology at the University of Massachusetts at Amherst. In 1994, he received his PhD from Lehigh University and prior to that was a practitioner in the public schools for 10 years. His research interests are in CBM and various forms of progress monitoring, research design, and data analysis that informs practice.

Randy G. Floyd is Associate Professor of Psychology at The University of Memphis. In 1999, he received his PhD from Indiana State University. His research interests include the structure, measurement, and correlates of cognitive abilities; the technical properties of early numeracy measures; and the process of professional publication.