School Psychology Review,
2015, Volume 44, No. 1, pp. 41–59

Using a Brief Preschool Early Numeracy Skills Screener to Identify Young Children With Mathematics Difficulties

David J. Purpura
Purdue University

Erin E. Reid
Erikson Institute

Michael D. Eiland
Wisconsin Center for Education Research

Arthur J. Baroody
University of Illinois at Urbana–Champaign and University of Denver

Abstract. A critical component in enhancing academic success is identifying children at risk of later academic difficulties. Although significant efforts have been devoted to design effective assessment processes in elementary school, fewer efforts (particularly for mathematics) have been made for preschool. The focus of this study was to design and evaluate a brief early numeracy skills screening tool. Measure development and validation occurred in a two-stage process with diverse and distinct samples. In the first stage, 393 preschool children were assessed on a battery of early numeracy tasks. By use of an item response theory framework, 24 items that spanned the ability continuum were selected for inclusion in the brief measure. In the second stage, 129 preschool children were assessed on the brief measure, the Test of Early Mathematics Ability–Third Edition, and two literacy measures. The data resulted in acceptable psychometric properties and strong diagnostic accuracy. Theoretical and practical implications are discussed.

Attaining basic mathematical competencies is a key factor in later academic and career success (Baroody, Lai, & Mix, 2006; National Mathematics Advisory Panel [NMAP], 2008). Children begin to develop individual differences in numeracy skills even by the preschool years (Berch, 2005; Stevenson et al., 1990), and these differences predict

This work was supported by grants from the Institute of Education Sciences, U.S. Department of Education (R305B04074 and R305B100017). The views expressed herein are solely those of the authors and have not been reviewed or cleared by the grantors.

Please address correspondence regarding this article to David J. Purpura, Purdue University, Human Development and Family Studies, 1202 W State St, Room 231, West Lafayette, IN 47907; e-mail: davidjamespurpura@gmail.com

Copyright 2015 by the National Association of School Psychologists, ISSN 0279-6015, eISSN 2372-966x

their later achievement (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Purpura, Baroody, & Lonigan, 2013). Moreover, such competencies are the foundation not only for advanced mathematical knowledge but also for achievement in academic areas such as science and engineering (Claessens & Engle, 2013; NMAP, 2008). With support, preschoolers can develop a wealth of early mathematical knowledge (Clements & Sarama, 2007). However, for teachers to identify which students are in need of more targeted support, it is necessary to use valid and reliable assessment tools.

The wide range of individual differences and the growing number of children at risk of later mathematics difficulties make the task of early mathematics instruction challenging for teachers (Berch, 2005). Preschool teachers must be able to identify quickly and accurately which children need additional evaluation and instruction and must be able to reliably measure their progress across the year. Unfortunately, teachers often do not have the time or training to formally test children's knowledge (Darling-Hammond, 2000). Importantly, teachers need a range of appropriate assessment tools to meet specific needs, including the ability to screen and monitor progress within their classrooms, so that they can plan whole-class, small-group, and individual instruction.

EXISTING MEASURES

There has been an increase in recent years in the development and validation of different types of brief mathematics measures for use in preschool. Efforts have been made to develop both measures of discrete skills and, to a lesser extent, brief broad-content measures of the continuum of skills. Discrete measures generally are fluency-based (Floyd, Hojnoski, & Key, 2006; Hojnoski, Silberglitt, & Floyd, 2009; Polignano & Hojnoski, 2012; VanDerHeyden, Broussard, & Cooley, 2006; VanDerHeyden, Broussard, Fabre, Stanley, Legendre, & Creppell, 2004) and non–fluency-based measures (Lei, Wu, DiPerna, & Morgan, 2009; Reid, Morgan, DiPerna, & Lei, 2006) that assess individual components and specific mathematical skills. Alternatively, brief broad-content measures are generally untimed and focus on multiple mathematical components. Broad content–focused measures can be administered one to three times per year and are used primarily to determine general ability levels and identify children at risk of later difficulties. Little work has been conducted in early mathematics to develop brief broad measures, but two brief versions of longer diagnostic measures have been developed that generally meet this framework: Child Math Assessment–Abbreviated (Klein & Starkey, 2000) and Brief Research-based Early Mathematics Assessment (Weiland et al., 2012).

LIMITATIONS OF EXISTING BRIEF MEASURES

The discrete skill measures have been shown to be good predictors of later mathematics performance and change in individual abilities over time (Hojnoski et al., 2009; VanDerHeyden et al., 2006). However, discrete skill measures have two key limitations that may affect their utility in discriminating between children at risk of later academic difficulties and those not at risk. First, mathematics fluency has been found to be distinct from untimed mathematics knowledge–based skills. Petrill et al. (2012) found that reading skills, particularly reading fluency, explained more variance in mathematical fluency than did untimed mathematics performance. This concern is further enhanced when evaluating the relations between fluency-based mathematics measures and nonmathematical measures. For example, VanDerHeyden et al. (2004) found that their fluency measures were about equally correlated with a preschool measure of mathematics as with a general school readiness measure. Similarly, Polignano and Hojnoski (2012) found that their fluency-based measures were highly correlated with nonmathematics tasks such as naming letters and colors. Essentially, fluency-based measures may not be measuring constructs that are distinctly mathematics.

The second issue related to discrete skill measures is that they target narrow areas of content (e.g., numeral identification, missing

number). In early mathematics, unlike in early literacy, no core deficit suggestive of early mathematics difficulties has been identified (Chiappe, 2005). Rather, mathematics skills, particularly in the early years, develop as a sequence of connected concepts and skills (NMAP, 2008) called a *learning trajectory* (Sarama & Clements, 2009). As such, Foegen, Jiban, and Deno (2007) indicated that a brief assessment tool that covered a broader range of content may actually be a better means by which to assess early mathematics skills in young children than a narrowly focused measure. With a brief broad-content screener, it may not be possible to include the full spectrum of skills that would be included in a lengthy learning trajectories assessment like the Tools for Early Assessment in Math (Clements, Sarama, & Wolfe, 2011) or the Test of Early Mathematics Ability–Third Edition (TEMA-3; Ginsburg & Baroody, 2003), but a brief measure that covers a range of content, albeit not every aspect of mathematics, would allow for the screening of children's performance on the content covered during the preschool years.

Although Foegen et al. (2007) recommended the broader-content approach, there is a significant dearth of available measures that offer the simplicity and ease of use that are needed for progress-monitoring measures. Two research groups have developed psychometrically strong brief forms of their full measures (Klein & Starkey, 2000; Weiland et al., 2012). Unfortunately, the practicality of using the briefer measures as screeners in preschool settings is limited. For example, not only do the measures take considerably more time than most progress-monitoring measures (approximately 15 min), but they also both use manipulatives, which are time-consuming to use, can be easily lost, and may make administering the test more complicated.

CURRENT STUDY

Ultimately, in the identification of students who need enhanced instruction, the goal is to minimize the number of children who are misidentified. It is possible that a brief broad-content screening tool would be better than discrete skill measures for identifying preschool children at risk of later difficulties (Foegen et al., 2007). However, it is also likely that a two-phase approach is preferred (Catts, Fey, Zhang, & Tomblin, 2001; Gilbert, Compton, Fuchs, & Fuchs, 2012) in which a short broad-content screener is used to distinguish between children who may need more enhanced instruction and those who are not at risk of later difficulties. This screener would then be followed by progress monitoring for those identified as at risk of later academic difficulties to identify the children most in need of additional instruction. To address either of these approaches, there is a critical need to develop a brief and easy-to-use broad mathematics screener that can be used across the preschool age range. Thus, the purpose of this study was to develop a brief measure of broad preschool mathematics skills and assess the utility of the resulting measure in identifying students as at risk of mathematics difficulties. The following goals guided the study:

1. Identify 20 to 25 easy-to-administer items that measure early numeracy skills across a range of ability and provide preliminary evidence of the reliability of this measure.
2. Identify if a ceiling rule can be included without reducing the reliability of the measure.
3. Determine if the brief measure exhibits acceptable convergent and discriminant validity.
4. Ensure the measure is sensitive to age-related differences in performance.
5. Determine the measure's utility in identifying children at risk of mathematics difficulties by identifying cutoff scores and evaluate if the ideal cutoff scores vary by age because children's mathematical knowledge develops rapidly during preschool (Ginsburg, Klein, & Starkey, 1998).

METHOD

The present study followed a two-stage development process. In Stage 1, the brief

measure was developed by selecting a subset of items from a broader pool of items. In Stage 2, assessment of reliability, validity, and screening utility was conducted. Because two separate samples and procedures were used for each stage, the Method and Results sections contain two subsections describing Stage 1 and Stage 2. However, because the two stages are logically sequential, the two Method subsections are presented first, followed by the two Results subsections.

Stage 1: Measure Development

The first stage in identifying preschool students who are at risk of mathematics difficulties is to develop a brief measure of broad skills. Thus, students were assessed using a broad pool of items similar to those used in previous research.

Participants

Data were collected using a convenience sample in 44 public and private preschools serving children from families with low to middle socioeconomic statuses. The 393 children who completed the testing were evenly split by sex (51.7% female) and approximately representative of the demographic characteristics of the local area (55.7% White, 33.8% Black, and 10.5% other race or ethnicity). Children ranged in age from 3.13 years to 5.98 years (M = 4.75 years, SD = 0.75 years), were primarily English speaking, and had no known developmental disorders based on teacher or school administrator report. Parental consent was obtained for each participating child.

Measures

Children were assessed on 25 measures of early numeracy skills (see Table 1 for descriptions of each task). The measures were developed and evaluated in our prior research (Purpura & Lonigan, 2013, 2015), and the tasks are representative of the range of skills assessed by other early numeracy measures (Clements, Sarama, & Liu, 2008; Ginsburg & Baroody, 2003; Griffin & Case, 1997; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Klein & Starkey, 2006; van de Rijt, Van Luit, & Pennings, 1999). These tasks were designed to assess the aspects of early numeracy skills deemed critical for success by the National Council of Teachers of Mathematics (NCTM, 2006) and the NMAP (2008). Furthermore, the skills and concepts measured are key developmental precursors to kindergarten mathematics success as noted in the Common Core State Standards (2010).

Procedure

Children were assessed on all 207 initial items; however, only the 143 items identified in prior work (Purpura & Lonigan, 2015) as being nonoverlapping (i.e., providing unique information) were included in the analyses. Assessments were conducted by individuals who either had completed or were working toward completion of a bachelor's degree. The assessors completed a 2- to 3-hour training session and were required to demonstrate proficiency on the assessment procedures before beginning data collection. Assessments occurred in the local preschools during noninstructional time in a quiet room designated by the individual preschool directors. Total testing time for each child was approximately 90 min. Assessments typically were conducted in three to four separate testing sessions.

Analytic Procedure

The primary focus of Stage 1 was to reduce the total number of items from 143 to approximately 20 to 25 items that varied in difficulty from relatively easy to relatively hard for preschool children through a three-step process. The analyses for doing so are described in the following sections.

Step 1. Item-level parameters were calculated using a two-parameter logistic (2-PL) item response theory (IRT) analysis in Mplus (Muthén & Muthén, 2012). IRT is a model-based method of latent trait measurement that relates the amount of an individual's latent ability to the probability of correctly responding to an item (Hambleton, Swaminathan, & Rogers, 1981). IRT allows researchers to select items based on item-level characteristics. The item-level characteristics, or parameters, in a 2-PL model that describe item functioning

Table 1. Description and Reliability for the Tasks From Which Items Were Selected

Skill	Description	No. of Items	α
Verbal counting	The child was asked to count as high as possible. After a mistake or when the child counted to 100, the task was stopped. The child was awarded one point each for correctly counting to 5, 10, 15, 20, 25, 30, 40, and 100.	8	—
Counting forward or backward[ab]	The examiner started a count sequence (either forward or backward) and, at a specified number, instructed the child to continue counting until told to stop.	6	.82
Counting error identification[cd]	The child was asked to identify correct or incorrect sequences of counting. Some of the count sequences were incorrect because of repetition of numbers, reversal of numbers in the counting sequence, skipping numbers, or skipping dots. If the child stated that the sequence was incorrect, the examiner asked the child what was wrong with the counting sequence. The child received one point for recognizing and identifying an error.	5	.84
One-to-one counting	The child was presented with a set of dots and was asked to count the set.	5	.79
Cardinality	In the context of the one-to-one counting task, the child was asked to indicate how many dots he or she had counted. The child was scored as answering correctly if he or she named the last number he or she counted, without re-counting the set.	3	.75
Resultative counting[a]	The child was asked to count a set of dots without touching the dots as he or she counted.	4	.68
Count a subset[abcdef]	The child was presented with a specific quantity of objects and was asked to count out a smaller set of objects from the larger set. In the second part of this task, which also had four items, the child was presented with a set of pictures of both dogs and cars and was asked how many of one set there were.	8	.82
Subitizing[b]	The child was briefly presented with a set and had to identify the total without counting.	7	.69
Estimation[d]	On the first two items, the child was shown a set of dots (e.g., 10 or 20) and was asked to estimate the number of dots on the page. A response was considered correct if the child provided a response within 25% of the exact answer. On the other three items, the child was presented with four sets of dots (10, 20, 50, and 100) and was asked to identify which was a specific number.	5	.49
Ordinality[abc]	The child was presented with a line of pictures and was asked to identify the *n*th picture.	5	.73

(Table 1 continues)

Table 1. Continued

Skill	Description	No. of Items	α
Relative size[def]	On the first two items of this task, the child was shown an array of five numbers (one number at the top of the page and four at the bottom of the page). He or she was asked to identify which of the numbers at the bottom of the page was numerically closest to the number at the top of the page. The second two items were just presented verbally.	4	.67
Number comparison[bcdef]	The child was asked to identify the largest or smallest number. Half the items were presented visually with Arabic numerals, and half the items were presented verbally.	6	.74
Set comparison[abef]	The child was presented with four sets of dots and was asked which had the most or fewest.	6	.77
Number order[bcdf]	The child was shown a number line and was asked to identify the number that was missing.	6	.87
Sequencing[b]	The child had to order three to five cards each with a set of dots of varying quantity.	4	.79
Set reproduction[ad]	For the first three items, the child matched his or her set of blocks to the experimenter's set. For the last three, the child identified which set (of four) was the same as a main set.	6	.63
Numeral identification	The child was presented with flashcards of nine numbers that ranged from 1–15. He or she was shown the flashcards one at a time and was asked, "What number is this?"	9	.90
Set to numerals[b]	On the first three items, the child matched a numeral to one of four sets of dots. On the last three items, he or she matched a set of dots to one of four numerals.	5	.80
Addition or subtraction with objects[bcd]	A set of discs was placed in a box; then a second set of discs was placed in (or removed from) the box. The child had to identify the total number in the box.	8	.72
Story problems	The child was presented verbally with basic addition and subtraction story problems.	7	.71
Initial equivalence[c]	Two empty boxes were placed on the table. The examiner placed a series of discs in the boxes one at a time. When the examiner had finished placing discs in each box, the child was asked if the boxes contained the same or a different number of discs.	6	.47
Two-set addition or subtraction[c]	After each question on the initial equivalence task, the examiner added discs to or subtracted discs from one of the boxes. The child was then asked if the boxes had the same or a different number of objects. The child was awarded one point as part of the two-set addition or subtraction score for correct responses on the second part of this task.	4	.41

(Table 1 continues)

Table 1. Continued

Skill	Description	No. of Items	α
Equivalent sets[ce]	The child was presented with a set of blocks (e.g., six) and a picture of a different quantity of objects (e.g., a picture of three tables). The child was instructed to divide the objects equally among the images so that all sets were equal.	5	.74
Number composition and decomposition[b]	The child was presented with a set of objects on the table and informed verbally of the quantity. The objects were hidden, and more objects were either added or subtracted from the initial set. The new set was presented, and the child was asked to identify how many objects were either added or subtracted from the initial set.	6	.74
Number combinations	The child was presented with a problem (e.g., 1 + 1) and was asked, "How much is __ [a + b]?"	5	.77

Note. Task consistent with items used on [a]Early Numeracy Test (van de Rijt, Van Luit, & Pennings, 1999); [b]Brief Research Based Early Mathematics Assessment; [c]Child Math Assessment–Abbreviated; [d]Number Sense Core battery (Jordan et al., 2007); [e]Test of Early Mathematics Abilities–Third Edition; and [f]Number Knowledge Test (Griffin & Case, 1997).

are referred to as the *difficulty parameter* and the *discrimination parameter*. The difficulty parameter measures the point along the ability spectrum at which a specific response option would be endorsed 50% of the time for an individual with a given ability. Items with high difficulty parameters require a greater amount of latent ability; hence, such items are more difficult to answer correctly than items with lower difficulty parameters. The discrimination parameter measures how well an item differentiates between individuals with latent abilities above and below the item's difficulty parameter. Other IRT models such as a one-parameter logistic (only the difficulty parameter is estimated) or three-parameter logistic (3-PL; an additional guessing parameter is estimated) exist; however, the 2-PL model was chosen over the one-parameter logistic model because selection of items with higher discrimination parameters (over items that had similar difficulty parameters but lower discrimination parameters) was desired to enhance the ability of the measure to reliably differentiate between individuals at different abilities. Furthermore, the 2-PL model was selected over the 3-PL model because the sample size for 3-PL models (>1,000) could not be practically justified given the small number of items that were multiple choice (<30%).

Step 2. Items that required manipulatives such as blocks or discs (e.g., addition and subtraction with objects, the first half of the set reproduction task) were removed from the next step of item selection. These items were removed from subsequent analyses because a key goal of this measure development process was to construct a measure that matched the discrete skill measures in their simplicity of administration. Items that required manipulatives generally took longer to administer. Furthermore, to ensure that the final measure was straightforward and quick to administer, tasks were removed from the next step of item selection if they (a) required additional steps to score (e.g., verbal counting, estimation); (b) required extended initial instructions and precise administration procedures that, if not followed exactly, would invalidate the item administration (e.g., counting forward or backward, counting error identification, resultative

counting, subitizing); or (c) were dependent on the administration of another item (e.g., cardinality—"how many"). These difficult-to-administer tasks were not desirable on a brief screener because their inclusion would potentially complicate the test administration. Although only items on certain tasks were used in the final selection process, all items were included in the Step 1 IRT analysis to ensure that item parameters were estimated on a broad conceptualization of numeracy skills. To ensure that removing these items did not alter the nature of the construct being measured, the total score of the brief measure was compared with a latent factor score that subsumed the common variance from all items in the initial item pool.

Step 3. The remaining items from Steps 1 and 2 were organized by item difficulty, and overlapping items were removed. The use of multiple items that provided overlapping information was not desired because it would result in a "double counting" of one difficulty level of item, which could inflate some children's total scores. For example, if the test had many items with identically low difficulty parameters (i.e., easy items), then the test scores would be inflated artificially at the low end of the scale (e.g., the sum of correct answers for five easy items is not equivalent to the sum of answers for five difficult items). However, if the test were constructed with several items that spanned the range of mathematical ability, none of which overlapped in their information, the test would be a uniform measure of mathematics ability across the range of the latent trait.

To determine which of two or more overlapping items were to be removed, two primary criteria were employed: (a) the item with the higher discrimination parameter was typically retained and (b) when both discrimination and difficulty parameters were comparable, items from tasks that were underrepresented in the already selected items were retained (e.g., if one of the overlapping items was a one-to-one counting item and the other item was a set comparison item, the item selected would depend on how many of each type of item had already been selected). A balanced approach to item selection was used to ensure that there were a breadth of items selected for the final measure. The overall goal of this item-reduction step was to reduce the number of total items to between 20 and 25 items while maintaining acceptable reliability over a broad range of latent abilities. The goal was to retain a test-level standard error of ideally less than .316 but not greater than .548 because standard errors of .316 and .548 are equivalent to classical test theory (CTT) internal consistencies of .90 and .70, respectively. In addition to IRT standard error scores, CTT reliability was calculated for the final measure.

Stage 2: Measure Validation and Utility as a Screener

The sample of children in Stage 2 participated in a larger intervention study. These data were from the pretest of that study.

Participants

Data were collected using a convenience sample in four preschool centers (two public and two private) serving children from families with low to middle socioeconomic statuses. The 129 children who completed all assessments were relatively evenly split by sex (58.1% female) and approximately representative of the demographic characteristics of the area (46.5% White, 31.0% Black, 14.0% Asian, and 8.5% Hispanic). Children ranged in age from 3.63 to 5.85 years (M = 4.79 years, SD = 0.49 years), were primarily English speaking, and had no known developmental disorders based on teacher or school administrator report. Parental consent was obtained for each participating child.

Measures

Each child was assessed with four measures, one of which was the newly devised brief screener. The measures are described in the following sections.

Preschool Early Numeracy Skills Screener–Brief Version. The 24 items selected in the previous stage were assessed together in one screening measure, the Pre-

school Early Numeracy Skills Screener–Brief Version (PENS-B). Items were ordered by difficulty, with the easiest items tested first. All 24 items were administered to all children.

Test of Early Mathematics Ability–Third Edition. The TEMA-3 (Ginsburg & Baroody, 2003) is a measure of informal and formal numeracy skills for children aged 3 years 0 months through 8 years 11 months. It is composed of 72 items and can be used as both a norm-referenced and diagnostic test. Administration time is typically around 30 to 40 min. Internal consistencies of the test and its alternate form are greater than 0.92 across ages included in this sample. Basal and ceiling rules were administered per the instruction manual.

Expressive One-Word Picture Vocabulary Test–Fourth Edition. The Expressive One-Word Picture Vocabulary Test–Fourth Edition (EOWPVT) (Martin & Brownell, 2011) was used to measure children's expressive vocabulary ability. In this task, children were shown a colored picture of an object(s) and asked, "What is this?," "What is this for?," or "What are these?" The EOWPVT has excellent reliability ($\alpha = 0.95–0.96$ for 3- to 5-year-old children; Martin & Brownell, 2011). Children received one point for each correct response. Basal and ceiling rules were administered per the instruction manual.

Get Ready to Read–Revised. The Get Ready to Read–Revised (GRTR) measure (Lonigan & Wilson, 2008) is a 25-item measure of print knowledge and phonological awareness. Children are administered all 25 multiple-choice items and earn one point for each correct response. Internal consistency for the normative sample in preschool was 0.88 (Lonigan & Wilson, 2008).

Procedures

As noted earlier, the sample of children in this stage participated in a larger intervention study; these data were from the pretest of that study. Assessments were conducted by individuals who had completed a bachelor's degree in education or psychology and either were working toward or had completed a doctoral degree. The assessors each completed a 1- to 2-hour training session on the PENS-B, EOWPVT, and GRTR measures and were required to demonstrate proficiency on the assessment procedures before beginning data collection. During the first 2 weeks of data collection, the second author observed each tester to ensure standardized administration procedures were followed for each measure, scored along with the tester, and provided feedback as needed. Assessors who administered the TEMA-3 all had significant prior training and experience with this measure. Assessments occurred in the local preschools in a quiet area designated by the individual preschool directors. Total testing time for the larger battery of assessments was approximately 60 to 90 min per child. Assessments typically were conducted in two to three separate testing sessions.

Analytic Procedure

Before assessment of the reliability and validity of the measure, two potential ceiling rules were identified (three or four incorrect in a row; Goal 2). A ceiling rule was desired for two key reasons. First, as items were ordered according to difficulty, once children incorrectly responded to multiple items in a row, the likelihood of correctly answering subsequent items was diminished significantly. Continued experience with items to which children would likely respond incorrectly could adversely affect other assessments given during the same testing period or their interest and motivation in future testing. Second, as children incorrectly responded to multiple items in a row, the likelihood that a subsequent correct response was due to chance increased. Including a ceiling rule would minimize chance correct responses and increase reliability of the test. The subsequent reliability and validity analyses all were conducted using all three ceiling rules (no ceiling, four in a row incorrect, and three in a row incorrect). The analyses for all three ceiling rules were conducted with the goal of selecting the small-

est ceiling rule that does not reduce the reliability of the measure.

To evaluate reliability and validity of the PENS-B, four key criteria were evaluated: sample-specific internal consistency, split-half reliability, concurrent convergent validity, and concurrent discriminant validity (Goal 3). Internal consistency was calculated using Cronbach's α. Split-half reliability was calculated by splitting the test by odd and even numbers and calculating the correlation between the two halves. Concurrent convergent validity was determined by calculating the correlation between the PENS-B and the TEMA-3. Correlations of $r > .50$ indicate significant convergent validity. Discriminant validity was calculated by comparing correlated correlations to show that the correlation between the PENS-B and the TEMA-3 was significantly higher than the correlation between the PENS-B and other nonmathematical variables (GRTR and EOWPVT). To assess the significance of discriminant validity, a process determined by Meng, Rosenthal, and Rubin (1992) was used. Correlations for convergent and discriminant validity were compared to determine if the two correlations were significantly different. If the convergent validity correlation was greater than .50 and was significantly higher than the discriminant validity correlation, this would be indicative of the measure having both convergent and discriminant validity.

To examine age-related differences in performance on the PENS-B and ensure that the psychometric properties of the measure were acceptable for all ages in preschool (Goal 4), children in both samples were combined ($N = 522$) and then divided into six age groups: 3 years 0 months to 3 years 5 months ($n = 28$), 3 years 6 months to 3 years 11 months ($n = 66$), 4 years 0 months to 4 years 5 months ($n = 74$), 4 years 6 months to 4 years 11 months ($n = 129$), 5 years 0 months to 5 years 5 months ($n = 135$), and 5 years 6 months to 5 years 11 months ($n = 90$). Means and effect size differences were compared across all age groups. Reliability, skewness, and kurtosis were also calculated for each age group.

To examine risk-status prediction (Goal 5) using the second sample ($n = 129$), receiver operating curves (ROC) were conducted for each age (3 year olds, 4 year olds, and 5 year olds) separately and the three age groups combined. The ROC analyses were used to identify a cutoff criterion score on the PENS-B that would maximize diagnostic accuracy for children at risk of later mathematics difficulties. Sensitivity (the proportion of students correctly classified as at risk) and specificity (the proportion of students correctly classified as not at risk) were used to identify ideal cutoff scores. Classification of risk status was based on a TEMA-3 score of 90 or less (i.e., the 25th percentile or lower). The goal in identifying a cutoff on the PENS-B was to maximize the number of children correctly classified as at risk of mathematics difficulties. The score at which the PENS-B had a sensitivity of .90 was selected. A sensitivity of .90 (and not the point at which sensitivity and specificity were equivalent) was selected because for an initial screening tool, it is better to overidentify children at risk of later difficulties and then remove false positives through subsequent assessment methods than to underidentify and miss children in need of further instruction. As such, positive predictive value (PPV; the likelihood that if the child fails the screener, he or she would also fail the diagnostic test) and negative predictive value (NPV; the likelihood that if a child passes the screener, he or she would go on to fail the diagnostic test) were also used as measures of classification accuracy. High PPV and low NPV are the goal; however, for an initial screening tool, it is critical to minimize the number of at-risk children missed by the screener (i.e., low NPV).

RESULTS

Stage 1: Initial Item Selection (Goal 1)

The first stage of the research was to create a brief measure of mathematics. To do this, item difficulty and discrimination parameters were calculated for the 143 potential items using a 2-PL IRT model in Mplus.

Step 1

Item difficulty and discrimination parameters were calculated using a 2-PL IRT model in Mplus. For the 143 initial items, difficulty parameters ranged from –3.52 to 5.30 and discrimination parameters ranged from 0.10 to 2.16. Because only a subset of the items were considered for inclusion in the final measure, the difficulty and discrimination parameters for each item are presented after Step 3, but only for the items considered for inclusion. However, it should be noted that, in general, the items with difficulty parameters below –2.00 and above 2.00 had lower discrimination parameters and were answered correctly by nearly all participants (for items $<$–2.00) or by nearly no students (for items $>$2.00).

Step 2

The tasks that included manipulatives were cardinality ("give me *n*"), set sequencing, set reproduction, addition or subtraction with objects, initial equivalence, two-set addition, fair sharing, and set composition or decomposition. Elimination of these tasks for inclusion in the final measure resulted in a decrease of 40 total items, from 143 to 103. Tasks that were difficult to administer or score and did not include manipulatives included verbal counting, counting forward or backward, identification of counting errors, cardinality ("how many"), resultative counting, subitizing, and estimation. Elimination of these tasks for inclusion in the final measure resulted in a decrease of 41 total items, from 103 to 62.

To ensure that the removal of these items as a whole did not significantly limit the ability to select well-performing items for the final item pool, the discrimination parameters for the retained items, the difficult-to-administer items, and the items with manipulatives were compared. Both the retained items, $t(111) = 4.61$, $p < .001$, and the difficult-to-administer items, $t(68) = 2.58$, $p = .012$, had significantly higher discrimination parameters than the items with manipulatives. These results suggest that, in general, using items without manipulatives would likely result in a measure that is better able to distinguish between individuals of differing abilities. One notable exception included the cardinality ("give me *n*") items. Those items generally had high discrimination parameters (1.28 to 1.91), but the inclusion of those items on the final measure would have resulted in the use of additional manipulatives. Discrimination parameters were not significantly different between the difficult-to-administer items and the retained items: $t(103) = 1.13$, $p = .260$. After elimination of the items with manipulatives and those difficult to administer or score, the difficulty parameters of the items retained for final item selection ranged from –1.52 to 1.63 and the discrimination parameters ranged from 0.32 to 1.87.

Step 3

Overall, 24 items from the 62-item pool were selected for inclusion in the final measure: 2 counting a subset, 3 one-to-one counting, 1 numeral comparison, 2 number order, 1 numeral identification, 1 ordinality, 2 relative size, 2 set comparison, 3 set to numerals, 3 story problems, and 4 number combinations. The final 24 items had difficulty parameters from –1.52 to 1.64, and the difference between consecutive items was approximately 0.10 to 0.15. In Table 2 the item parameters of the 24 final items are presented from least to most difficult.

To ensure that minimal information was lost when removing items that used manipulatives or were difficult to administer, the total score of the final 24 items was correlated with a latent factor score that subsumed the common variance from all original 143 items. The two scores were highly correlated ($r = .94$, $p < .001$). This finding indicates that the brief measure functioned quite similarly to using the broader measure and that little information regarding broad numeracy skills was lost by removing sets of items.

Preliminary Reliability Evaluation

Reliability was evaluated through both CTT and IRT reliability methods. Internal consistency was high ($\alpha = .90$). However, α

Table 2. Item Parameters for Final Items Included in Preschool Early Numeracy Skills Test (n = 393)

Type of Task	Item	Discrimination	Difficulty
One-to-one counting	Count 3 dots	1.57	−1.52
Counting a subset	Count 3 pictures from a larger set	1.68	−1.47
Set comparison—most	Identify 8 dots as largest of 4 sets	0.92	−1.35
Numeral identification	Identify the numeral 1	1.32	−1.26
Set to numerals	Connect the numeral 1 to 1 dot	1.41	−1.11
One-to-one counting	Count 6 dots	1.19	−0.96
Set comparison—most	Identify 3 dots as largest of 4 sets	1.06	−0.81
Set to numerals	Connect the numeral 3 to 3 dots	1.66	−0.67
One-to-one counting	Count 11 dots	0.93	−0.50
Set to numerals	Connect the numeral 5 to 5 dots	1.05	−0.38
Number order	Identify number before 5	1.43	−0.25
Relative size	Identify number closest to 4 from 4 options	0.80	−0.14
Story problems	1 + 1 =	0.71	0.00
Number order	Identify number after 9	1.87	0.12
Number comparison—most	Identify numeral 8 as largest of 4 numerals	0.76	0.28
Story problems	1–1 =	0.94	0.38
Relative size	Identify number closest to 9 from 4 options	0.97	0.53
Story problems	4–1 =	0.70	0.77
Number combinations	1 + 1 =	0.73	0.90
Counting a subset	Count 20 pictures from of a larger set	0.85	1.05
Number combinations	2 + 2 =	0.90	1.16
Ordinality	Identify eighth object	0.86	1.31
Number combinations	0 + 2 =	0.32	1.49
Number combinations	1 + 3 =	0.67	1.64

scores provide only test-level reliability and are sample dependent. IRT standard error scores provide reliability information across the spectrum of ability and thus are sample independent. The final task had standard errors of less than .316 for theta scores between –1.60 and 0.40 and less than .548 for theta scores between –2.40 and 1.90, indicating that the test was reliable across a wide range of ability.

Stage 2: Utility as a Screener

The next stage in the research was to examine if the new measures could be used to screen preschool students. The first step toward that goal was to develop a ceiling rule, examine the reliability and convergent validity of the data for this group because screeners can only have utility if they meet basic psychometric standards, and assess the diagnostic accuracy of the data. Descriptive statistics and correlations for all measures are presented in Table 3.

Selection of a Ceiling Rule (Goal 2)

The goal in selecting a ceiling rule was to reduce overall testing time but maintain evidence of reliability and validity. There were limited differences on subsequent analyses for the three different ceiling rules. Using the three-in-a-row ceiling rule reduced testing for 78% of children and by an average of 7.32 items (from 24 items down to 16.68 items). Furthermore, 31% of participants had their total testing reduced by over 50% using this ceiling rule. Because this ceiling rule significantly decreased the amount of testing while maintaining evidence of reliability and validity, only the results using the three-in-a-row ceiling rule are presented.

Table 3. Descriptive Statistics and Correlations Between Variables Used in Measure Validation

	Descriptive Statistics					Correlations			
	M	*SD*	Range	Skewness	Kurtosis	PENS-B	TEMA-3	EOWPVT	GRTR
PENS-B	11.88	6.23	0–24	0.13	–0.95	—	.73	.60	.64
TEMA-3	15.28	10.57	0–68	1.36	3.82	.78	—	.58	.63
EOWPVT	59.05	19.50	14–116	0.26	−0.12	.61	.59	—	.66
GRTR	16.43	5.48	3–25	−0.39	−0.64	.70	.70	.67	—

Note. $n = 129$. All correlations were significant at $p < .001$. Correlations below the diagonal are zero-order correlations, and those above the diagonal are partial correlations controlling for age. EOWPVT = Expressive One-Word Picture Vocabulary Test–Fourth Edition; GRTR = Get Ready to Read; PENS-B = Preschool Early Numeracy Skills Test–Brief; TEMA-3 = Test of Early Mathematics Abilities–Third Edition.

Preliminary Analyses

Descriptive statistics and correlations for all measures are presented in Table 3. All variables were normally distributed. No sex differences were found in performance on any measure. All measures were significantly correlated.

Reliability and Validity (Goal 3)

The PENS-B exhibited high internal consistency ($\alpha = .93$) in the validation sample. Removal of any items would not have increased reliability because all items had item total correlations greater than 0.29 (median item total correlation = .57). Split-half reliability was also high ($r = .90$).

Zero-order correlations between the TEMA-3 and PENS-B showed that the PENS-B has strong convergent validity ($r = .78$). However, because age may have also affected the relation between the two measures, partial correlations between the measures were calculated accounting for age. Results were similar to the zero-order correlations and indicated that the measures were highly correlated even when accounting for age ($r = .73$).

Correlations between the math measures and language or literacy measures were calculated using both zero-order and partial correlations (accounting for age) and are presented in Table 3. Although the partial correlations between the PENS-B and the language or literacy measures were high (GRTR, $r = .64$; EOWPVT, $r = .60$), they were comparable to the correlations between the TEMA-3 and the language or literacy measures (GRTR, $r = .63$; EOWPVT, $r = .58$). To test if the PENS-B was significantly more correlated with the TEMA-3 than it was with the language or literacy measures, a test of correlated correlations was conducted. Partial correlations were used in these analyses. The correlation between the PENS-B and the TEMA-3 was significantly higher than the correlations between the PENS-B and the EOWPVT ($z = 2.73$, $p = .003$) and GRTR ($z = 1.97$, $p = .024$) measures, showing evidence of both convergent and discriminant validity for the PENS-B.

Age-Related Evaluation (Goal 4)

There were significant differences in mean performance across all age groups ($ps < .01$) with the exception of the comparison between the two youngest age groups: $t(92) = 1.25$, $p = .214$, Hedge's $g = 0.28$. Furthermore, the screener exhibited strong reliability and acceptable levels of skewness and kurtosis across all age groups (see Table 4).

Risk-Status Prediction (Goal 5)

In the validation sample, 43% of children were classified as at risk of mathematics difficulties based on TEMA-3 scores of 90 or below. Analyses for all three age groups re-

Table 4. Descriptive Statistics of PENS-B by Age Group

Age Group	*n*	Range	*M*	*SD*	Skewness	Kurtosis	Reliability	ES Dif
3 years 0 months to 3 years 5 months	28	0–12	3.82	3.17	0.57	−0.09	.80	0.28
3 years 6 months to 3 years 11 months	66	0–18	4.86	3.89	0.89	0.88	.87	0.73*
4 years 0 months to 4 years 5 months	74	0–21	8.18	4.99	0.41	–0.52	.90	0.59*
4 years 6 months to 4 years 11 months	129	0–24	11.57	6.07	−0.10	−0.86	.92	0.48*
5 years 0 months to 5 years 5 months	135	1–23	14.36	5.41	−0.47	−0.49	.92	0.46*
5 years 6 months to 5 years 11 months	90	2–24	16.86	4.93	−0.71	0.18	.88	—

Note. $N = 522$. ES Dif = Hedge's *g* effect size difference (and significance test) between listed age group and next older age group; PENS-B = Preschool Early Numeracy Skills Test–Brief.
*$p < .01$.

sulted in specificities of .71 or above. Cutoffs for each age group logically increased with age, with PENS-B scores of 7 for 3 year olds, 10 for 4 year olds, and 15 for 5 year olds. Although the sample of 3 year olds was relatively small ($n = 10$), classification was perfect. By use of these three cutoff scores, the PENS-B resulted in an overall correct classification accuracy of 82.2% with strong PPVs and NPVs across the three ages (of the 23 misclassified children, only 5 were false negatives and 18 were false positives). Results are summarized in Table 5.

DISCUSSION

In this study, a brief screening measure of preschool early numeracy skills (the PENS-B) was developed and its utility for identifying preschool students as at risk of mathematics difficulties was determined. The items on the PENS-B broadly assess many of the key domains identified by the NCTM (2006) and NMAP (2008) as critically important for young children's numeracy development. The PENS-B takes less than 5 min to administer and could be well suited as an initial screening measure in preschool classrooms because it exhibited strong risk-status classification.

Age-Related Differences and Risk Status

Children's mathematical knowledge undergoes rapid and dramatic changes across the preschool years (Ginsburg et al., 1998). The concurrent age-related analyses in this study show that the PENS-B is sensitive to these differences. Though not statistically significant, there was a small effect size difference

Table 5. Risk-Status Prediction Analyses for the Three Age Groups

	n	Sensitivity	Specificity	PPV	NPV	Optimal Cut Score[a]	Children at Risk, %[b]
Full sample	129	.89	.67	.68	.11	12	43
5 year olds	51	.88	.80	.82	.13	15	51
4 year olds	68	.91	.71	.62	.06	10	34
3 year olds	10	1.00	1.00	1.00	.00	7	70

Note. NPV = negative predictive value; PPV = positive predictive value.
[a]Scores equal to or below the cut score on the Preschool Early Numeracy Skills Test–Brief are highly predictive of children also scoring equal to or below a standard score of 90 on the Test of Early Math Abilities–Third Edition (TEMA-3). [b]Percent of children who scored equal to or below a standard score of 90 on the TEMA-3.

(Hedge's $g = 0.28$) found between the two youngest age groups. This effect size, coupled with the small sample size for the youngest age group (3 year olds), may suggest that in subsequent studies, statistically significant and important differences may be found in these age groups. Critically, the age-related differences found in this study stand in contrast to other studies using discrete skill measures (e.g., Floyd et al., 2006), in which it was found that it was not possible to distinguish the performance of 3 and 4 year olds. These findings support the assertion made by Foegen et al. (2007) that a broad-content screener may be more useful than discrete skill measures for the assessment of mathematics with younger children.

One of the key features of this measure is its utility for assessing risk status. Using a strict cutoff that did not factor in age resulted in an overall classification accuracy of 76%. However, this rate increased to 82% when considering age. Importantly, of the 18% of children misclassified, only 5 (4% of the total sample) were false negatives, meaning they were classified as at risk of mathematics difficulties based on the criterion (TEMA-3) but were not identified as such on the PENS-B. Furthermore, the age-based cutoffs provide a relatively simple screening mechanism that aligns with early mathematical learning trajectories (Sarama & Clements, 2009). For example, Items 1 to 7 on the PENS-B are all basic counting (sets of no more than six), set comparison ("Which set has the most dots?"), or numeral identification items—skills that 3 year olds are expected to attain. The next three items include more advanced set counting (11 dots) and two items that combine numeral recognition and cardinality skills (connecting a set to a numeral for numbers under 5)—skills that 4 year olds are expected to attain. Finally, the next five items primarily involve number ordering and magnitude, as well as basic story problems—skills typically attained by 5 year olds.

Some caution should be applied to using these as strict age-based cutoffs. First, more narrow age ranges (e.g., 3- or 6-month age bands instead of 12-month age bands) may be more appropriate and might result in greater classification accuracy. The relatively small sample in this second dataset prevented a more targeted classification approach. In fact, in the second sample, there were only ten 3 year olds and they were all 3.50 to 3.99 years old. Further evaluation and validation of the cutoff criteria are needed in a broader sample. Second, these risk-status cutoffs are intended to be general guidelines. Depending on a school's population and resources to provide further instruction, the cutoffs may need to be modified to best fit the needs of the school (VanDerHeyden, 2011). Third, the cut score and evaluation of its accuracy were conducted on the same sample, which may have inflated the accuracy. Further independent evaluation of these cut scores is warranted. Finally, the benefit of using a screener (such as the PENS-B) or progress-monitoring measures needs to be empirically evaluated. Although the PENS-B generally performed well in terms of classification accuracy indices, there were still some students (18 of 129, or 14% of the sample) who were overclassified as being at risk of mathematics difficulties. Ultimately, a two-stage screening process that combines an initial static screening and a subsequent progress-monitoring approach may be the best method for truly identifying those children at risk of later difficulties (Catts et al., 2001; Gilbert et al., 2012); however, for early mathematics, this is a question in need of further research.

Relation of Numeracy Skills to Nonmathematics Domains

The validity analyses provide evidence that the PENS-B has both convergent and discriminant validity. The PENS-B is significantly more related to another measure of numeracy (TEMA-3) than it is to measures of early literacy skills (EOWPVT and GRTR), suggesting that it is measuring a distinct component of early achievement (numeracy) compared with the early literacy measures. These findings are in contrast to discrete skill measures that found nearly identical correlations with measures of mathematics and nonmath-

ematics domains (Polignano & Hojnoski, 2012; VanDerHeyden et al., 2004). However, it is also clear in the current study and in previous literature (LeFevre et al., 2010; Purpura, Hume, Sims, & Lonigan, 2011) that early mathematics skills have a strong language- and print-based component. For example, many early mathematics terms and applications, such as the words *more* and *less* or *combine* and *take away*, are effectively language terms, and language skills seem to be broadly related to most early numeracy skills (Purpura & Ganley, 2014). Similarly, letter and numeral knowledge skills are highly related (Austin, Blevins-Knabe, Ota, Rowe, & Lindauer, 2011; Piasta, Purpura, & Wagner, 2010) likely because they share many underlying code-related features (Brizuela, 2004). Interestingly, Dirks, Spyer, van Lieshout, and de Sonneville (2008) found that it was more common for children to have difficulties in both mathematics and reading than in either one alone. Given the high relation between the two domains and the joint-risk probability, it may be beneficial to include measures of literacy in a risk assessment for mathematics difficulties. Including mathematical and nonmathematical assessments (such as literacy) in a risk-assessment battery may enhance risk-status identification and enable researchers and practitioners to identify children who otherwise would not be properly identified.

Usability

In developing the PENS-B, practical issues related to usability were addressed. Specifically, the length of administration was a key consideration in designing the measure because a screening tool needs to be able to be administered to a large number of students quickly. As a result, the PENS-B takes under 5 min to administer after the application of the ceiling rule; administration of the screener takes approximately 33% less time than without the ceiling rule (an average testing time of $<$3.5 min). This short administration time will allow teachers to assess their whole class in a relatively brief amount of time ($<$75 min for a typical class of 15 students) and then conduct more targeted and detailed assessment of children most in need. This practical benefit also carries over to research. Because academic domains are often interrelated, researchers in nonmathematics domains (e.g., literacy and executive functioning) are often interested in using mathematics measures to understand cross-domain relations. However, the extensive time needed to administer many broad measures may limit such opportunities. A brief mathematics measure such as the PENS-B will provide an opportunity to expand research efforts to cross multiple domains without the need to devote considerable additional time and resources to assessments.

Limitations

Although the findings presented herein are of interest to both researchers and practitioners, several limitations should be noted. First, although prior evidence (Purpura & Lonigan, 2015) showed that there was no differential item functioning (DIF) either for sex or for race/ethnicity across any of these items, other variables such as socioeconomic status (SES) could result in DIF. DIF could not be calculated on family SES in these samples because family demographic information was not collected. In future studies, SES-related DIF and a more in-depth exploration regarding race-based DIF (particularly in reference to English language learners) should be examined. Second, even though the age-based analyses within this study indicate that there are significant differences across most age groups, these data were concurrent in nature. The utility of the PENS-B for being sensitive to intra-individual change cannot be ascertained from the current study. Future work using the PENS-B to show growth over time needs to be conducted. Furthermore, assessments of short- and long-term test–retest reliability, as well as predictive validity studies, need to be conducted. Third, this measure covers a relatively broad range of numeracy domains that have been identified as key skills and concepts in preschool. However, there is still a need to construct and validate a separate brief measure of early geometry because numeracy and ge-

ometry have been found to be distinct domains in preschool (Wolfe, Clements, & Sarama, 2011). Finally, although strong psychometric properties of this measure have been shown, ultimately, the success of any assessment tool rests in its adoption and use by teachers. If teachers are not willing to use a measure (whether it is because the measure is complicated to administer or they do not perceive that it is appropriate for that age group or construct), the practical application of the measure in schools will be limited. Further research is needed to assess (and potentially enhance) the social validity of this measure, as well as to determine the extent of training needed for teachers to administer the measure.

REFERENCES

Aunola, K., Leskinen, E., Lerkkanen, M., & Nurmi, J. (2004). Developmental dynamics of math performances from preschool to grade 2. *Journal of Educational Psychology, 96,* 699–713.

Austin, A. M. B., Blevins-Knabe, B., Ota, C., Rowe, T., & Lindauer, S. L. K. (2011). Mediators of preschoolers' early mathematics concepts. *Early Child Development and Care, 181,* 1181–1198.

Baroody, A. J., Lai, M., & Mix, K. S. (2006). Development of young children's early number and operation sense and its implications for early childhood education. In Spodek, B. & Saracho, O. N. (Eds.), *Handbook of research on the education of young children* (2nd ed.) (pp. 187–221). Mahwah, NJ: Lawrence Erlbaum Associates.

Berch, D. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities, 38,* 333–339.

Brizuela, B. M. (2004). *Mathematical development in young children: Exploring notations.* New York, NY: Teachers College Press.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32,* 38–50.

Chiappe, P. (2005). How reading research can inform mathematics difficulties: The search for the core deficit. *Journal of Learning Disabilities, 38,* 313–317.

Claessens, A., & Engel, M. (2013). How important is it where you start? Early mathematics and later school success. *Teachers College Record, 115,* 1–29.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education, 38,* 136–163.

Clements, D. H., Sarama, J., & Liu, X. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Maths Assessment. *Educational Psychology, 28,* 457–482.

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for Early Assessment in Mathematics.* Columbus, OH: McGraw-Hill Education.

Common Core State Standards. (2010). *Common core state standards: Preparing America's students for college and career.* Retrieved from http://www.corestandards.org/

Darling-Hammond, L. (2000). *Solving the dilemmas of teacher supply, demand, and standards: How we can ensure a competent, caring, and qualified teacher for every child.* New York, NY: National Commission on Teaching & America's Future.

Dirks, E., Spyer, G., van Lieshout, E. C. D. M., & de Sonneville, L. (2008). Prevalence of combined reading and arithmetic disabilities. *Journal of Learning Disabilities, 41,* 460–473.

Floyd, R. G., Hojnoski, R., & Key, J. (2006). Preliminary evidence of the technical adequacy of the preschool numeracy indicators. *School Psychology Review, 35,* 627–644.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measuring in mathematics: A review of the literature. *The Journal of Special Education, 41,* 121–139.

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for a four-step screening system. *Assessment for Effective Intervention, 38,* 6–14.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). Austin, TX: Pro-Ed.

Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In Williams, D., Sigel, I. E., & Renninger, K. (Eds.), *Child psychology in practice* (pp. 401–476). Hoboken, NJ: John Wiley & Sons.

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 2,* 1–49.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1981). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hojnoski, R. L., Silberglitt, B., & Floyd, R. G. (2009). Sensitivity to growth over time of the preschool numeracy indicators with a sample of preschoolers in Head Start. *School Psychology Review, 38,* 402–418.

Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22,* 36–46.

Klein, A., & Starkey, P. (2000). *Child Math Assessment–Abbreviated.* Berkeley, CA: Author.

Klein, A., & Starkey, P. (2006). *Child Math Assessment.* Berkeley, CA: Author.

LeFevre, J., Fast, L., Skwarchuk, S., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81,* 1753–1767.

Lei, P., Wu, Q., DiPerna, J. C., & Morgan, P. L. (2009). Developing short forms of the EARLI numeracy measures: Comparison of item selection methods. *Educational and Psychological Measurement, 69,* 825–842.

Lonigan, C. J., & Wilson, S. B. (2008). *Report on the revised Get Ready to Read! screening tool: Psychometrics and normative information* [Technical report].

New York, NY: National Center for Learning Disabilities.

Martin, N. A., & Brownell, R. (2011). *Expressive one-word picture vocabulary test manual* (4th ed.). Novato, CA: Academic Therapy Publications.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Muthén, L. K., & Muthén, B. O. (2012). Mplus 7.0 [Computer program]. Los Angeles, CA: Muthén & Muthén.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Petrill, S., Logan, J., Hart, S., Vincent, P., Thompson, L., Kovas, Y., & Plomin, R. (2012). Math fluency is etiologically distinct from untimed math performance, decoding fluency, and untimed reading performance: Evidence from a twin study. *Journal of Learning Disabilities, 54,* 371–381.

Piasta, S. B., Purpura, D. J., & Wagner, R. (2010). Fostering alphabet knowledge development: A comparison of two instructional approaches. *Reading and Writing, 23,* 607–626.

Polignano, J. C., & Hojnoski, R. L. (2012). Preliminary evidence of the technical adequacy of additional curriculum-based measures for preschool mathematics. *Assessment for Effective Education, 37,* 70–83.

Purpura, D. J., Baroody, A. J., & Lonigan, C. J. (2013). The transition from informal to formal mathematical knowledge: Mediation by numeral knowledge. *Journal of Educational Psychology, 105,* 453–464.

Purpura, D. J., & Ganley, C. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology, 122,* 104–121.

Purpura, D. J., Hume, L., Sims, D., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology, 110,* 647–658.

Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations in preschool. *American Educational Research Journal, 50,* 178–209.

Purpura, D. J., & Lonigan, C. J. (2015). Early numeracy assessment: The development of the preschool early numeracy scales. *Early Education and Development, 26,* 286–313. doi:10.1080/10409289.2015.991084

Reid, E. E., Morgan, P. L., DiPerna, J. C., & Lei, P. (2006). Development of measures to assess young children's early academic skills: Preliminary findings from a Head Start-university partnership. *Insights on Learning Disabilities, 3,* 25–38.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* New York, NY: Routledge.

Stevenson, H. W., Lee, S., Chen, C., Lummis, M., Stigler, J., Fan, L., & Ge, F. (1990). Mathematics achievement of children in China and the United States. *Child Development, 61,* 1053–1066.

van de Rijt, B. A. M., Van Luit, J. E. H., & Pennings, A. H. (1999). The construction of the Utrecht Early Mathematical Competence Scales. *Educational and Psychological Measurement, 59,* 289–309.

VanDerHeyden, A., Broussard, C., Fabre, M., Stanley, J., Legendre, J., & Creppell, R. (2004). Development and validation of curriculum-based measures of math performance for preschool children. *Journal of Early Intervention, 27,* 27–41.

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children, 77,* 335–350.

VanDerHeyden, A. M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. *Journal of School Psychology, 44,* 533–553.

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. C., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of a short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology, 32,* 311–333.

Wolfe, C. B., Clements, D. H., & Sarama, J. (2011, March). *A factorial invariance analysis of early mathematics assessment with prekindergarteners.* Poster presented at the Biennial Meeting of the Society for Research in Child Development, Montreal, Quebec.

Date Received: February 11, 2014
Date Accepted: October 22, 2014
Associate Editor: Amanda VanDerHeyden ■

David J. Purpura is an assistant professor of human development and family studies at Purdue University. His research focuses on improving preschool children's mathematics and literacy acquisition through early individualized assessment and intervention. He also focuses on understanding the connections between mathematics and nonmathematical domains such as literacy and cognitive skills and how those domains affect the acquisition of mathematical knowledge.

Erin E. Reid, PhD, is a postdoctoral research fellow for the Early Math Collaborative at the Erikson Institute. Her research focuses on the learning and teaching of mathematical skills and concepts in early childhood. She is also interested in developing measures to assess the skills, behaviors, attitudes, and practices that lead to mathematics competence in young children and the effective teaching of early mathematics.

Michael D. Eiland is an Institute of Education Sciences postdoctoral research fellow with the Wisconsin Center for Education Research. He received his doctorate in curriculum and instruction from the University of Illinois at Urbana–Champaign. His research interests include early childhood and elementary mathematics education.

Arthur J. Baroody is currently a professor emeritus of curriculum and instruction at the University of Illinois at Urbana–Champaign and a senior research fellow at the Morgridge College of Education, University of Denver. He received his PhD in educational and developmental psychology from Cornell University in 1979. His research focuses on the development of number, counting, and arithmetic concepts and skills from preschool to Grade 2.