School Psychology Review,
2010, Volume 39, No. 2, pp. 181–195

SPECIAL TOPIC

Validating a Number Sense Screening Tool for Use in Kindergarten and First Grade: Prediction of Mathematics Proficiency in Third Grade

Nancy C. Jordan and Joseph Glutting
University of Delaware

Chaitanya Ramineni
Educational Testing Service

Marley W. Watkins
Arizona State University

Abstract. Using a longitudinal design, children were given a brief number sense screener (NSB) screener ($N = 204$) over six time points, from the beginning of kindergarten to the middle of first grade. The NSB is based on research showing the importance of number competence (number, number relations, and number operations) for success in mathematics. Children's mathematics achievement on a validated high-stakes state test was measured 3 years later, at the end of third grade. Test–retest reliability estimates were obtained for the NSB. Two criterion groups were then formed on the basis of the third-grade achievement test (children who met and who did not meet mathematics standards). Diagnostic validity analyses for the NSB were completed using repeated measures analyses of variance and receiver operator curve analyses. Results from all analyses revealed that scores on the NSB in kindergarten and first grade predicted mathematics proficiency in third grade. Areas under the receiver operator curve indicated that the NSB has high diagnostic accuracy (areas under the receiver operator curve = 0.78–0.88). Findings suggest that kindergarten and first-grade performance on the NSB is meaningful for predicting which children experience later mathematics difficulties.

This work is supported by a grant from the National Institute of Child Health and Human Development (R01HD036672). The authors thank the participating children and teachers for their extremely generous cooperation.

Correspondence regarding this article should be addressed to Nancy C. Jordan, Department of Education, 211C Willard Hall, University of Delaware, Newark, DE 19716; E-mail: njordan@udel.edu

Mathematics education is rapidly becoming a top priority among U.S. policy makers because proficiency in mathematics is essential to success in the disciplines of science, technology, engineering, and mathematics and for competiveness in the global workforce. Poor mathematics achievement is widespread in U.S. schools, especially among economically disadvantaged, minority populations. Disturbingly, substantial mathematics disparities exist between middle- and low-income children before they enter school (Jordan & Levine, 2009; National Research Council, 2009). Recent research indicates the importance of early number competence, or *number sense*, for setting children's learning trajectories in mathematics throughout elementary school (Duncan et al., 2008; Jordan, Kaplan, Ramineni, & Locuniak, 2009). Number sense refers to the understanding of whole numbers, number operations, and number relations (Malofeeva, Day, Saco, Young, & Ciancio, 2004; National Research Council, 2009). Number sense allows children to connect mathematical principles with procedures (Gersten, Jordan, & Flojo, 2005).

Most children enter school with number sense that is relevant to learning formal mathematics (National Research Council, 2009). Even in the first year of life, humans are sensitive to numerical and related spatial representations (e.g., Antell & Keating, 1983; Cordes & Brannon, 2008; Wynn, 1992). Infants have precise representations of small sets of objects and approximate representations of large sets (Feigenson & Carey, 2003). These primary abilities appear to develop without much verbal input or instruction (Berch, 2005; Dehaene, 1997; Feigneson, Dehaene, & Spelke, 2004). Preverbal number knowledge is shared by children from different cultures and cognitive abilities (Gordon, 2004; Pica, Lerner, Izard, & Dehaene, 2004) and, arguably, provides a foundation for acquiring secondary symbolic number competencies related to counting, comparing, and operating on sets. Knowledge of the symbolic number system is influenced by the input a child receives and can be taught successfully in preschool and kindergarten (Ginsburg, Lee, & Boyd, 2008; Siegler, 2009). Symbolic number sense is secondary to primary preverbal number knowledge but intermediate to the formal mathematics that is taught in school (Jordan & Levine, 2009).

Counting knowledge is critical for extending quantitative understanding beyond small numbers (Baroody, 1987; Baroody, Lai, & Mix. 2006; Ginsburg, 1989). Children begin to say the count words soon after they learn to talk (Fuson, 1988). At first, they might use the count words to label small quantities of 3 or less or recite the count list; later, they might use the count words for counting objects in a set. Before kindergarten, most children internalize key counting principles (Gelman & Gallistel, 1978)—that is, each item can be counted only once, the count words must be used in a stable consistent order, and the final number in the count indicates how many items are in the set.

Children as young as 4 years of age also learn to discriminate between and among quantities (Case & Griffin, 1990; Griffin, 2002, 2004). For example, they can tell which of two piles of objects has more or less. By 6 years of age, most children integrate these quantitative sensitivities with their counting knowledge to form a mental number line (Siegler & Booth, 2004). They eventually understand that numbers later in the count list have larger quantities than earlier quantities (N, $N + 1$, $[N + 1] + 1$, and so on; Le Corre & Carey, 2006) and that, for example, 4 is bigger than 3 and that 2 is smaller than 5 (Griffin, 2004).

Counting and quantity discrimination help children perform addition and subtraction calculations. Although preschoolers have limited success with addition and subtraction story problems ("Mike had 2 pennies. Barb gave him 1 more penny. How many pennies does he have now?") and number combinations ("How much is 2 and 1?"), many can solve "nonverbal" problems with object representations (e.g., The child is shown 2 objects that are then hidden with a cover. One more object is slid under the cover and the child must indicate that 3 objects are now under the cover; Ginsburg & Russell, 1981; Levine, Jor-

dan & Huttenlocher, 1992). Preschool performance on nonverbal calculations is associated with kindergarten performance on story problems and number combinations, suggesting that nonverbal mental models may underpin calculations with number words (Levine et al., 1992; Huttenlocher, Jordan, & Levine, 1993).

Mathematics learning difficulties and disabilities appear to have their roots in weak number sense related to whole numbers, number relations, and number operations, as opposed to more general cognitive deficits (Landerl, Bevan, & Butterworth, 2004; Malofeeva et al., 2004; Mazzocco & Thompson, 2005). *Dyscalculia*, a severe form of mathematics disability, is characterized by domain-specific impairments in number more than general impairments related to language, memory, or spatial knowledge. Poor number sense is reflected by weak counting procedures, slow fact retrieval, and inaccurate computation (Geary, Hamson, & Hoard, 2000; Jordan, Hanich, & Kaplan, 2003a, 2003b).

Predictability of Number Sense

Number sense can be reliably measured in young children and is predictive of later mathematics achievement outcomes (Clarke & Shinn, 2004). In short-term longitudinal studies (fall to spring of the school year), it has been shown that numeracy indicators of oral counting, quantity discrimination, quantity identification, number identification, and naming missing numbers in a sequence are moderate to strong predictors of mathematics achievement (correlations ranging from .33 to .79, with quantity discrimination being particularly strong (Clark & Shinn, 2004; Lembke & Foegen, 2009). Methe, Hintze, and Floyd (2008) found fluency measures related to ordinal position and number recognition fluency (given in the fall of kindergarten) are strongly predictive of performance on a mathematics test in the spring of kindergarten (r = .58 and .72, respectively). Moreover, preschool numeracy measures given in the spring of the 4-year-old year are correlated moderately with kindergarten numeracy measures obtained during the winter of kindergarten (VanDerHeyden, Broussard, & Cooley, 2006).

Using a "core" number sense battery designed for use in kindergarten and first grade, Jordan and colleagues (Jordan, Kaplan, Olah, & Locuniak, 2006; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Jordan et al., 2009) examined children's number competence from the beginning of kindergarten to the middle of first grade in relation to their mathematics achievement and growth between the end of first grade and the end of third grade using the Woodcock-Johnson Tests of Achievement—III (WJ-III; McGrew, Schrank, & Woodcock, 2007). The battery explicitly incorporated key research findings related to number, number relations, and number operations into the number sense measure (National Research Council, 2009), and the approach was to follow children over multiple years. Overall, it was shown that children develop foundational number sense before first grade that supports their learning of more complex mathematics. Specifically: (a) Kindergarten number sense predicted rate of growth in mathematics achievement between first and third grades as well as achievement level through third grade. (b) The relatively poor first- through third-grade mathematics achievement of low-income children was *mediated* through their weak number sense in kindergarten. (c) Kindergarten number sense related to addition and subtraction operations was most predictive of later mathematics achievement. Jordan et al. (2009) observed that " if children leave kindergarten with weak number competencies, especially with respect to operational knowledge and skills, they will enter first grade at a disadvantage and may never catch up to children who started kindergarten with good number competencies" (p.864), which may lead to a "cascade of mathematics failure in school" (p. 865).

Present Study

The present study provides additional evidence of the potential importance of number sense screening for early identification of mathematics difficulties. The core number

sense battery developed by Jordan and colleagues (e.g., Jordan et al., 2006; Jordan et al., 2007; Locuniak & Jordan, 2008; Jordan, Glutting, & Ramineni, 2010) has a relatively long administration time (about 35 min), making it less practical for use as a screening tool in schools. Jordan, Glutting et al. (2008) developed an abbreviated but reliable screening version using Rasch item analyses. This shortened measure is referred to as the number sense brief (NSB). Using the NSB, the present investigation compared number sense performance of children who met versus those who failed to meet proficiency standards on a high-stakes state mathematics test in third grade. Number sense was examined longitudinally over six time points, from the beginning of kindergarten to the middle of first grade.

The present investigation expands our understanding in several ways. First, it establishes the diagnostic accuracy of the NSB using conventional procedures (*t* tests and repeated-measures analyses of variance) for evaluating test-score validity. Second, the study then employs receiver operating characteristic (ROC) curve analyses to assess diagnostic accuracy. Statistically significant group differences have been the traditional benchmark for determining whether a test has discriminant and/or criterion-related validity (cf. Watkins, 2009 for a review). That is, the mean score of individuals with a particular disorder (e.g., low achievement or learning disabilities) is compared to the mean score of individuals without the problem (e.g., normal achievement). Statistically significant group differences are then interpreted as evidence that the test is diagnostically effective.

Although mean score differences indicate that *groups* can be discriminated, this conventional validity approach cannot be uncritically extended to conclude that mean group differences are distinctive enough to differentiate among *individuals*. As noted by Elwood (1993), "significance alone does *not* reflect the size of the group differences nor does it imply the test can discriminate subjects with sufficient accuracy for clinical use" (p. 409, original italics). Little attention has been paid to the overlap in score distributions between regular and exceptional groups, although its importance has been known for decades (Meehl & Rosen, 1955). In other words, group mean differences are necessary but not sufficient for making accurate decisions about individuals because they do not take into account the overlap in score distributions between groups (Weiner, 2003).

In contrast, diagnostic utility statistics directly examine the individual aspects of decision making (Streiner, 2003; Watkins, 2005). Sensitivity and specificity statistics are among the most common measures used to evaluate diagnostic utility (Kessel & Zimmerman, 1993). *Sensitivity* is the proportion of individuals with a disorder (e.g., individuals with low achievement or learning disabilities) who are correctly identified by a positive test finding. *Specificity* is the proportion of individuals without the disability who are correctly identified by a negative test finding. Although useful, sensitivity and specificity can be adversely affected by cut scores (the score on a test used to differentiate success from failure; Swets, Dawes, & Monahan, 2000). Typically, the two statistics move in opposite directions when a new cut score is employed. Thus, if the cut score is varied, then sensitivity and specificity values will also change (Meehl & Rosen, 1955).

The ROC curve is an alternative diagnostic-utility statistic. By systematically using all possible cut scores of a test and plotting the true-positive rate (i.e., sensitivity) against the false-positive rate (i.e., 1—specificity) for each cut score, diagnostic validity can be displayed for the full range of the test's scores (McFall & Treat, 1999; Swets et al., 2000). Unlike the sensitivity and specificity statistics, the ROC procedure is not dependent on the value of a specific cut score (Rey, Morris-Yates, & Stanislaw, 1992). The ROC also is independent of assumptions about the normality of a test's score distribution (Hanley & McNeil, 1982). Consequently, a ROC is "the state-of-the-art method" for describing the diagnostic accuracy of a test (Weinstein, Obuchowski, & Lieber, 2005, p. 16) and is "recognized widely as the most meaningful approach to quantify the accuracy of diagnostic

information and diagnostic decisions" (Metz & Pan, 1999, p. 1).

Visually, the test's true-positive rate is plotted on the *X* axis and the false-positive rate on the *Y* axis of a graph while systematically moving the test's cut score across its full range of values. A 45° diagonal line is drawn on the graph, which is the "random ROC" and reflects a test with zero discriminating power. The more clearly a test is able to discriminate between individuals with and without a disorder, the farther its ROC curve deviates toward the upper left corner of the graph.

The accuracy of a ROC can be quantified by calculating its AUC, or area under the curve (Henderson, 1993). Chance diagnostic accuracy corresponds to an AUC of 0.50, signifying that the true-positive rates and false-positive rates are equal across all possible cut scores and that the test provides no discrimination (Swets, 1988). On the other hand, perfect diagnostic accuracy corresponds to an AUC of 1.00.

The AUC is a measure of effect size. An AUC of 0.556 represents a small effect size, 0.639 indicates a medium effect size, and 0.714 and above denotes a large effect size (Rice & Harris, 2005). In addition, AUC can be viewed as a measure of clinical significance, where values of <0.70 are poor, 0.70–0.79 are fair, 0.80–0.89 are good, and 0.90–1.00 are excellent (Cicchetti, 2001). The AUC also is easy to understand. For instance, if one person is randomly selected from the nondisabled population and another person is randomly selected from the disabled population, the AUC is the probability of distinguishing between those two individuals with the test (McFall & Treat, 1999). Another advantage of ROCs is that the plotting of sensitivity and specificity values can be used to establish optimal cut points (Fletcher, Fletcher, & Wagner, 1996). It also is possible to compare AUCs and statistically determine whether the diagnostic accuracy of two correlated forms is equal (e.g., compare the validity of a short form to see if it is as accurate as a test's longer form; Hanley & McNeil, 1983).

The current study uses effect sizes from the ROC to estimate the magnitude of the NSB's accuracy at six different time points, beginning in kindergarten and proceeding to the middle of first -grade. Further, ROCs will be compared between the NSB and the original, longer version of the test at each time point to determine whether diagnostic accuracy was equivalent between the NSB and its longer version. In addition, at each time point, ROCs will be employed to establish optimal cut points that maximize the short form's diagnostic sensitivity (identification of children who are true positives) and its diagnostic specificity (identification of children who are true negatives). Overabundance of false positives can lead to wasted resources whereas high false negatives may deprive at-risk children of intervention (Fuchs, Fuchs, Compton, Bryant, Hamlett, & Seethaler, 2007).

Method

Participants

Participants were drawn from a 4-year longitudinal investigation of children's mathematics development (Jordan et al., 2006). They attended the same public school district in northern Delaware. All kindergartners from six schools were invited to participate in the study. We received Institutional Review Board approved informed consent from approximately 66% of the children. There were 378 children who started the study at the beginning of kindergarten and 204 who remained at the end of third grade. Participant attrition was from children moving out of the school district (typically right after kindergarten), rather than withdrawal from the study or absence on the day of testing. A logistical regression analysis (Jordan et al., 2009) revealed that although gender and age did not predict the odds of being absent from the study in third grade, low-income and minority children, respectively, were about 1.2 times more likely to be absent from the study than middle-income and nonminority children. In third grade, 52% of the children were boys, 45% were minority (63% African American, 26% Hispanic, and 11% Asian), and 23% came from low-income families. Income status was determined by participation in the free or re-

Table 1
Demographic Information for Participants by Group

Demographic Variable	Met DSTP Math Standards (n = 172)	Did Not Meet DSTP Math Standards (n = 32)
Gender		
Male	56%	31%
Female	44%	69%
Income[a]		
Low income	17%	59%
Middle income	83%	41%
Mean age	66 months (4 months)	66 months (4 months)

Note. DSTP = Delaware State Testing Program. Standard deviations in parentheses.
[a]Low income was determined by eligibility for the school district's free or reduced-price lunch program.

duced-price lunch program in school, and most low-income children resided in urban neighborhoods. Participant demographics are summarized in Table 1. All children were taught with the same mathematics curriculum (i.e., Math Trailblazers; Teaching Integrated Math and Science Curriculum, 2004) in kindergarten. All but 15 of the children continued with Math Trailblazers through third grade. These 15 students transferred to schools in the same district that used Investigations in Number, Data, and Space (Teacher Education Research Center, 1998). Both curricula were used in the district because they have similar content and are aligned to state standards.

Measures

NSB screen. The NSB has 33 items and is an untimed measure that takes approximately 15 min to administer. The items assess counting knowledge and principles (e.g., set enumeration, knowledge of the count sequence to at least 10, and principles of one-to-one correspondence, cardinality, and stable order); number recognition (e.g., the ability to name written symbols such as 13, 37, and 82); number knowledge (e.g., What number comes right after 7? Which number is bigger, 5 or 4?); nonverbal addition/subtraction calculations (e.g., The examiner places 2 chips in front of the child and then covers the chips. The examiner then puts out another chip and hides it under the same cover. The child indicates how many chips are now under the cover, either by putting out the appropriate number of chips form his or her own pile or by stating the answer); addition/subtraction story problems (e.g., orally presented as "Jill has 2 pennies. Jim gives her 1 more penny. How many pennies does Jill have now?"); and addition/subtraction number combinations (orally presented as "How much is 2 + 1?"). The measure is internally consistent, with a coefficient alpha of at least .80 at each time point in kindergarten and first grade (Jordan, Glutting et al., 2008). Each item was scored 0 (incorrect) or 1 (correct) with a total raw score of 33. Scoring is objective and thus there were no issues with interscorer reliability. The complete set of items for the NSB can be found in Jordan et al. (2010).

Jordan, Glutting et al. (2008) demonstrated that performance on the NSB at the beginning of first grade made a unique and meaningful contribution to the variance in WJ-III (McGrew et al., 2007) mathematics achievement at the end of both first and third grades, and the NSB did so after controlling for age and general cognitive abilities (i.e., language, spatial reasoning, and working memory). There were medium to large effect sizes for all analyses (Jordan et al., 2010). Strikingly, the predictability of the NSB was as strong for third-grade mathematics achieve-

Table 2
Test–Retest Reliability Coefficients

Time of Administration	Time of Administration					
	September K	November K	February K	April K	November Grade 1	February Grade 1
September K	—	.81	.80	.78	.69	.61
November K		—	.82	.81	.70	.61
February K			—	.86	.77	.70
April K				—	.81	.75
November Grade 1					—	.80
February Grade 1						—

ment as it was for first-grade mathematics achievement. This finding was unexpected because, whereas items on first-grade mathematics achievement tests are closely allied with those on the NSB, items on the third-grade tests are more varied (e.g., operations with rationale as well as whole numbers, multiplication and division as well as addition and subtraction) and less directly connected to what was measured on the NSB. The NSB also predicted achievement level on WJ-III subtests of written calculation and applied problem solving. Although performance on the NSB in kindergarten has a convergent predictive association with mathematics achievement at the end of third grade (r = .63), it has divergent predictive association with reading achievement (r = .29), suggesting further that the measure is uniquely related to mathematics (Jordan, Glutting et al., 2008). Such a pattern of high convergent correlations in conjunction with lower divergent correlations supports inferences that the NSB possess strong construct validity (Gregory, 2007; Reynolds, Livingston, & Wilson, 2006).

As a preliminary matter to the current investigation, test–retest reliability was investigated for the NSB. Table 2 presents test–retest reliability coefficients across the six time periods employed in the current study. As expected, stability coefficients are higher for shorter intervals. Reliabilities ranged from .61 to .86. Twelve of the 15 reliability coefficients were at, or above, the .70 criterion recommended in certain assessment textbooks (e.g., Gregory, 2007; Reynolds et al., 2006). Three coefficients dipped below the .70 criterion. However, this occurred only when the testing period exceeded 1 year, and the finding points to the need for annual retesting with the NSB.

Delaware Student Testing Program in Mathematics. Mathematics achievement was assessed with the third-grade version of The Delaware Student Testing Program (DSTP) in Mathematics (Delaware Department of Education, 2008). The DSTP measures concepts and procedures in accordance with Delaware mathematics standards (i.e., numeric reasoning, algebraic reasoning, geometric reasoning, and quantitative reasoning). It has strong internal reliability (.93) and has established cut scores for meeting state standards and for performing below state standards. The test's cut points and content were fully validated by a panel of experts (Delaware Department of Education, 2008). The DSTP in third grade is highly correlated (r = .77, $p <$ .01) with scores on the WJ-III—Mathematics (McGrew et al., 2007), indicating strong criterion validity (Jordan et al., 2009). For the present study, the DSTP mathematics outcome measure was used in categorical form (1 = met standards, 0 = did not meet standards). We collapsed the original five performance levels to two levels to simplify the measurement scale. The performance levels of 3 (*Meets the Standard*), 4 (*Exceeds the Stan-*

Table 3
Means and Standard Deviations by Group Across Six Time Periods for Number Sense Brief

Time	Met DSTP Math Standards[a]	Failed to Meet DSTP Math Standards[b]
September of kindergarten	17.3 (4.7)	11.7 (2.7)
November of kindergarten	19.2 (4.3)	15.2 (3.4)
February of kindergarten	20.5 (4.4)	15.1 (3.1)
April of kindergarten	21.8 (4.6)	16.4 (3.3)
November of first grade	24.1 (4.4)	18.7 (4.4)
February of first grade	26.2 (3.8)	21.3 (4.3)

Note. DSTP = Delaware State Testing Program. Values outside of parentheses represent means. Values within parentheses identify standard deviations.
[a]$N = 172$.
[b]$N = 32$.

dard), and 5 (*Distinguished Performance*) were transformed to a 1 on the categorical scale to represent meeting the DSTP standards, whereas the remaining two lower performance levels 1 (*Well below the Standard*) and 2 (*Below the Standard*) were transformed to a 0 on the categorical scale to denote failure to meet the standards on DSTP in mathematics. Participant demographics by group are presented in Table 1.

Procedure

The NSB was given to children individually in school by one of several trained graduate or undergraduate research assistants. It was administered in September, November, February, and April of kindergarten, and in November and February of first grade. (Students were originally given the longer version of the NSB, as described in Jordan et al., 2007; for the present study, the scoring was shortened to determine whether the NSB would perform as well as the longer version.) The DSTP was group administered by school personnel in April of third grade. No interventions were provided on the basis of the NSB screening.

Results

Table 3 presents means and standard deviations for scores from the NSB across the six kindergarten and first-grade time periods for children who met and did not meet DSTP mathematics standards in third grade. Figure 1 provides a visual representation by plotting mean scores across time and it does so separately by group (i.e., children who met standards vs. those who did not meet standards). Data were subsequently evaluated using a traditional, repeated-measures analysis of variance. Mauchly's test indicated that the assumption of sphericity was violated ($\chi^2 = 42.982$, $df = 14$, $p = .001$). Consequently, degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.895$). Results revealed that the main effect for time was significant [$F = 104.98$, df (4.476, 559.54), $p = .001$]. Partial eta squared (η^2) is one of the most popular measures of effect size in repeated-measures analyses of variance (Tabachnick & Fidell, 2001). Part of the reason for its popularity is that Cohen (1988) provided guidelines for interpreting small ($\eta^2 = 0.01$), medium ($\eta^2 = 0.09$), and large ($\eta^2 = 0.25$) effects. The obtained outcome represented a large effect size for time (i.e., $\eta^2 = 0.46$). Post hoc comparisons further demonstrated that the time trend was best described by a linear function [$F = 287.58$, df (1, 125), $p = .001$]. The significant main effect for time adds to the test's construct validity because the obtained raw scores increases on the NSB conform to appro-

priate developmental (age) changes expected for an achievement measure (Gregory, 2007; Reynolds et al., 2006).

As anticipated, results from the repeated-measures analysis of variance also showed a statistically significant main effect for group [$F = 37.96$, df (1, 125), $p = .001$]. The main effect for group revealed that children who met the DSTP math standard at the end of third grade consistently obtained higher NSB scores across time than those who did not meet the math standard. This difference represented a medium to large effect size (i.e., $\eta^2 = 0.233$). Lastly, as also anticipated, the group by time interaction was not significant [respectively, $F = 1.01$, df (7.030, 6.994), $p = .409$.

All six ROC curves lay markedly above and to the left of the diagonal, reflecting strong associations between all six administrations of the NSB and poor performance on the mathematics DSTP in third grade. More important, Table 4 provides statistics to accompany the ROC analysis. The second column in Table 4 (overall p values) provides significance levels. NSB scores at all six time points accurately discriminated between children who met and failed to meet third-grade DSTP mathematics standards. A more important set of values is located in the third column of Table 4. This column supplies AUC values for each time the NSB was administered. Every obtained AUC exceeded the critical value suggested for a large effect size (i.e., 0.714; Rice & Harris, 2005) and five of six exhibited good levels clinical significance (Cicchetti, 2001). Consequently, the NSB displayed high and meaningful levels of diagnostic accuracy. The last column of Table 4 compares AUCs between the NSB and the longer version of the NWB (Hanley & McNeil, 1983). Results showed that scores from the NSB and the longer version were equally accurate in predicting which

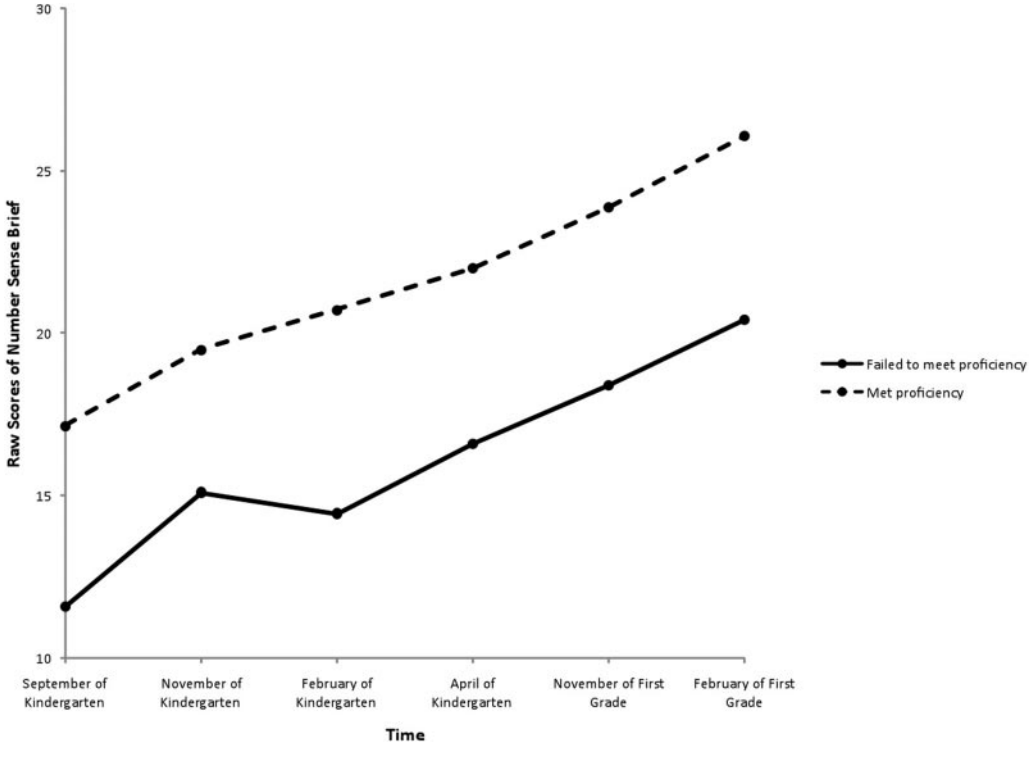

Figure 1. Number Sense Brief raw scores plotted separately across time according to whether students met, or failed to meet, proficiency on the mathematics portion of the Delaware State Testing Program.

Table 4
Diagnostic Utility Statistics for the Number Sense Brief

When NSB Was Administered	Overall *p* Values	AUC	Statistical Significance Level (*p* value) of AUC Comparisons Between the NSB and the Longer Version
September of kindergarten	.001	0.80	.02[a]
November of kindergarten	.001	0.81	.35
February of kindergarten	.001	0.86	.66
April of kindergarten	.001	0.80	.32
November of first grade	.001	0.78	.22
February of first grade	.001	0.88	.01[a]

Note. NSB = number sense brief; AUC = area under the curve.
[a]Difference favors NSB.

children would and would not meet mathematics standards in third grade.

Optimal cut scores were determined using two methods. The first method employed Swets and Picket's (1982) *d* statistic, where

$$d = \sqrt{(1 - sensitivity)^2 + (1 - specificity)^2}$$

The optimal cut score is defined as the point on a test's score continuum where the value of *d* is minimal (Yovanoff & Squires, 2006). The left side of Table 5 presents optimal cut scores using the *d* statistic. It also presents sensitivity and specificity values associated with each cut score. Results show that the sensitivities and specificities were high at each time point, thereby meeting the guidelines of the American Academy of Pediatrics (2001) for good diagnostic tests. The two values were also comparable at each interval and thereby approached optimal balancing (see Prigerson et al., 1999, on the importance of equal balancing of sensitivity and specificity values). At the same time, it is important to distinguish between "diagnosis" and "screening." The *d* statistic was designed to identify optimal *diagnostic* cut scores. On the other hand, the NSB is a screening measure. For screening measures, particularly those employed in a prevention model, examiners might want to err on the side of identifying all the children in need of preventative interventions. In such instances, it would be appropriate to sacrifice specificity in order to maximize sensitivity. Consequently, the right side of Table 5 presents optimal cut scores designed to maximize the NSB's sensitivity. Cut scores for the second method are based on sensitivity values $\geq .85$.

A comparison of *d*-based versus sensitivity-based cut scores reveals that the sensitivity-based cut scores were always lower. Therefore, sensitivity-based cut scores will identify more children who need further testing. Such an outcome is beneficial for a screening model where the goal is to maximize the identification of children who need help. In sum, outcomes from validity analyses revealed that kindergarten and first-grade performance on the NSB is meaningful for predicting which children will show mathematics weaknesses in third grade.

Discussion

Children's early number competencies were assessed longitudinally with a number sense brief screener (NSB) from the beginning of kindergarten through the middle of first grade. Students' mathematics proficiency was later assessed (at the end of third grade) with a high-stakes state test. The findings confirmed that the NSB has a high level of pre-

Table 5
ROC Cutoff Scores and Classification Statistics Across Six Number Sense Brief Time Points

	d-Based Optimal ROC[a]			Sensitivity-Based Optimal ROC[b]		
Time	Cut Score	Sensitivity	Specificity	Cut Score	Sensitivity	Specificity
September of kindergarten	15	.73	.85	13	.86	.65
November of kindergarten	18	.71	.85	15	.88	.35
February of kindergarten	16	.86	.75	16	.86	.75
April of kindergarten	20	.70	.85	17	.87	.50
November of first grade	22	.74	.70	20	.87	.60
February of first grade	24	.78	.75	22	.89	.55

Note. ROC = receiver operating characteristic.
[a]Optimal cut score based on Swets and Picket (1982) formula for *d*, where $d = \sqrt{(1 - sensitivity)^2 + (1 - specificity)^2}$.
[b]Optimal cut score based on maximizing sensitivity values ≥ .85.

dictive validity and diagnostic accuracy. In particular, there were large and statistically meaningful NSB differences between children who met and failed to meet third-grade mathematics standards. Moreover, low scores on the NSB result in decisions that are useful in predicting which children will fail to meet third-grade mathematics standards. AUCs measured by ROC analyses provide direct evidence of the NSB's high diagnostic accuracy (0.78 to 0.88).

These findings reveal that the NSB is a highly effective screen for young children who are likely to develop mathematics difficulties. They are in keeping with other studies suggesting that number sense is a powerful predictor of later mathematics outcomes (Mazzocco & Thompson, 2005). It has also been shown that number sense makes contributions to later mathematics outcomes, over and above general cognitive abilities, reading skill, and income status (Jordan et al., 2010; Locuniak & Jordan, 2008). Weaknesses in symbolic number sense, related to counting, number relationships, and basic operations, appear to underpin most mathematics learning difficulties (e.g., Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007; Gersten et al., 2005; Landerl et al., 2004).

To date, most studies have been concerned primarily with establishing a theoretical connection between number sense and mathematics outcomes, a connection that is highly compelling (Jordan et al., 2009). The present investigation indicates that the NSB, based explicitly on research models of early number development, has good potential for use by school psychologists in school or other clinical settings. The NSB items, by design, are closely tied to most U.S. mathematics curricula (National Council for Teachers of Mathematics, 2004) and thus identify potential intervention targets for kindergarten and early first grade. For example, children need to learn that numbers later in the count list have larger quantities and that numbers themselves have magnitudes (e.g., 4 is one more than 3 and one less than 5; Jordan & Levine, 2009). These understandings can bolster children's learning of addition and subtraction combinations. In keeping with other curriculum-based numeracy measures (e.g., Methe et al., 2008), the NSB shows a strong ability to predict long-term risk (through third grade) from the very beginning of kindergarten.

The NSB is relevant to response to intervention (RTI) service delivery models. The 2004 reauthorization of the Individuals with Disabilities Educational Improvement Act (Public Law 108–446) allows states to use response to intervention for identifying students with learning disabilities as an alter-

native to models emphasizing discrepancies between a student's IQ and achievement. Response to intervention is a multitiered prevention system for increasing student achievement (Fuchs et al., 2007). Critical components of response to intervention are as follows: (a) identify students at risk for poor achievement early, (b) provide evidence-based interventions, (c) monitor progress, (d) adjust instruction according to students' responsiveness, and (e) identify students with true learning disabilities rather than difficulties related to ineffective instruction (Fletcher & Vaughn, 2009; National Center on Response to Intervention, 2007). The goal of response to intervention is to intertwine assessment with academic programming. There is particular urgency for developing reliable screening tools in mathematics, which are scarcer than are those for reading (Gersten et al., 2005). The NSB has strong potential for accurately identifying students who are at risk for mathematics difficulties and who might need additional preventative instruction. Although false positives and negatives can never be eliminated entirely by a single measure, we were able to establish cut points on the NSB on six occasions spanning kindergarten and first grade with good sensitivity (identification of true positives) and specificity (identification of true negatives). The optimal cut scores for the six kindergarten and first-grade time points (see Table 5) could provide useful information for monitoring progress during this age period.

Several limitations, however, should be considered before putting the NSB into practice. First, the data are based on a sample from a single geographic location in the United States (i.e., mid-Atlantic region). Although the DSTP outcome measure is fully validated (Delaware Department of Education, 2008) and our sample was relatively diverse in terms of socioeconomic status and ethnicity, the present findings should be considered preliminary pending wider replication. Second, the NSB has not yet been studied to monitor progress made as a result of an intervention or some kind of alternative instruction. However, there is mounting evidence that number competencies as early as preschool can be improved through targeted intervention (Baroody, Eiland, & Thompson, 2009; Ramani & Siegler, 2008), and the NSB should be useful in this respect. It also would be worthwhile to test 1-month practice effects (Gregory 2007; Reynolds et al., 2006) because the NSB was given on multiple occasions. Another step would be to develop of alternate forms of the NSB for repeated use.

Although the NSB has a relatively short administration time, it may be possible to identify an even smaller subset of items that also have adequate predictability. For instance, Mazzocco and Thompson (2005) found that a composite of only four items from a math test was almost as predictive of which children would be diagnosed as learning disabled in second and third grades as a larger battery of tests. In addition to accuracy, response speed and especially strategy use on the number tasks deserve further consideration (Methe et al., 2008). Geary, Bailey, and Hoard (2009) found that the speed and accuracy with which children identify numbers and quantities of sets of objects that add up to a cardinal value represented by a number (e.g., the Arabic number 4 and one object makes 5) is highly predictive of mathematics learning disabilities in third grade. Moreover, the counting strategies children use to derive sums or differences (e.g., counting on from an addend to find a solution, using fingers) might add to the test's predictability and would have particular relevance for instructional planning (Baroody, Bajwa, & Eiland, 2009). Finally, the disproportionate attrition by minority and income status could reduce generalization of the third-grade findings. Although we could not determine why some children left the school district, the data suggest relatively high mobility among low-income families. Because high mobility is a risk factor for poor educational outcomes (Rumberger & Larson, 1998), one could argue that the findings from the present investigation do not adequately represent a group of high-risk children.

In sum, early number sense is an important predictor of elementary school mathematics difficulties. The NSB, used in the present study, is a research-based tool with good pre-

dictive validity. State-of the-art ROC analyses revealed high diagnostic accuracy for identifying children who could have benefited from early assistance in mathematics. Performance on the NSB also provides guidance for developing and validating early interventions in number competence. Whether gains in early number competence lead to sustained gains in mathematics achievement remains an open question. We are currently testing a kindergarten intervention, based on our model, in a randomized controlled study.

References

American Academy of Pediatrics. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics, 108,* 192–196.

Antell, S., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development, 54,* 695–701.

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education, 18*, 141–157.

Baroody, A. J., Bajwa, N. P., & Eiland, M. (2009). Why can't Johnny remember the basic facts? *Developmental Disabilities Research Reviews, 15,* 69–79.

Baroody. A. J., Eiland, M., & Thompson, B. (2009). Fostering at-risk preschoolers' number sense. *Early Education and Development, 20,* 80–128.

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

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

Case, R., & Griffin, S. (1990). Child cognitive development: The role of central conceptual structures in the development of scientific and social thought. In E. A. Hauert (Ed.), *Developmental psychology: Cognitive, perceptuo-motor, and neurological perspectives* (pp. 193–230). North-Holland: Elsevier.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23,* 695–700.

Cordes, S., & Brannon, E. M. (2008). Quantitative competencies in infancy. *Developmental Science, 11,* 803–808.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Delaware Department of Education. (2008). *Delaware state testing program technical report—2007.* Dover, DE: Author.

Duncan, G. J., Dowsett, C. J., Classens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review, 13*, 409–419.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science, 6,* 568–584.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *TRENDS in Cognitive Sciences, 8*, 307–314.

Fletcher, R. H., Fletcher, S. W., & Wagner, E. H. (1996). *Clinical epidemiology: The essentials* (3rd ed.). Baltimore: Williams & Wilkins.

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives, 3,* 30–37.

Fuchs, L. S., Fuchs, D., Compton, D., Bryant, J. D., Hamlett, C. L., & Seethhaler, P. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73,* 311–330.

Fuson, K. C. (1988). *Children's counting and concepts of number.* New York: Springer-Verlag.

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. *Journal of Psychoeducational Assessment, 27,* 265–279.

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology, 77,* 236–263.

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. *Child Development, 78,* 1343–1359.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38,* 293–304.

Ginsburg, H. P. (1989). *Children's arithmetic.* Austin, TX: PRO-ED.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report, 22*, 3–22.

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, Serial No. 193), 1–69.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306,* 496–499.

Gregory, R. J. (2007). *Psychological testing: History, principles, and applications* (5th ed.). Boston: Allyn & Bacon.

Griffin, S. (2004). Building number sense with Number Worlds: A mathematics program for young children. *Early Childhood Research Quarterly, 19,* 173–180.

Griffin, S. (2002). The development of math competence in the preschool and early school years: Cognitive foundations and instructional strategies. In J. M. Roher (Ed.), *Mathematical cognition* (pp. 1–32). Greenwich, CT: Information Age Publishing.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operator characteristic curves derived from the same cases. *Radiology, 148,* 839–843.

Henderson, A. R. (1993). Assessing test accuracy and its clinical consequences: A primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry, 30,* 521–539.

Huttenlocher, J., Jordan, N. C., & Levine, S. C. (1994). A mental model for early arithmetic. *Journal of Experimental Psychology: General, 123,* 284–296.

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). A number sense assessment tool for identifying children at risk for mathematical difficulties. In A. Dowker (Ed.), *Mathematical difficulties: Psychology and intervention* (pp. 45–58). San Diego, CA: Academic Press.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences, 20,* 82–88.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development, 74*, 834–850.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology, 85,* 103–119.

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.

Jordan, N. C., Kaplan, D., Olah, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77,* 153–175.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2008). Development of number combination skill in the early school years: When do fingers help? *Developmental Science, 11,* 662–668.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45,* 850–867.

Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews, 15,* 60–68.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

Landerl, K., Bevan, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition, 93,* 99–125.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105,* 395–438.

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice*, *24,* 12–20.

Levine, S. C., Jordan, N. C., & Huttenlocher, J. (1992). Development of calculation abilities in young children. *Journal of Experimental Child Psychology, 53,* 72–103.

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities, 41,* 451–459.

Malofeeva, E., Day, J., Saco, X., Young, L., & Ciancio, D. (2004). Construction and evaluation of a number sense test with Head Start children. *Journal of Educational Psychology, 96*, 648–659.

Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20,* 142–155.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Woodcock-Johnson III Normative Update.* Rolling Meadows, IL: Riverside Publishing.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Methe, S. A., Hintze, J. M., & Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review, 37,* 359–373.

Metz, C. E., & Pan, X. (1999). "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology, 43,* 1–33.

National Center on Response to Intervention. (2007). What is RTI? Retrieved June 8, 2009, from http://www.rti4success.org

National Council of Teachers of Mathematics. (2007). *Second handbook of research on mathematics teaching and learning.* Washington, DC: Author.

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity.* Committee on Early Childhood Mathematics, C. T. Cross, T. Woods, & H. Schweingruber (Eds.). Washington, DC: The National Academies Press.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian Indigene group. *Science,* 499–503.

Prigerson, H. G., Shear, M. K., Jacobs, S. C., Reynolds, C. F., Maciejewski, P. K., Davidson, J. R. T., et al. (1999). Consensus criteria for traumatic grief: A preliminary empirical test. *British Journal of Psychiatry, 174,* 67–73.

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development, 79,* 375–394.

Rey, J. M., Morris-Yates, A., & Stanislaw, H. (1992). Measuring the accuracy of diagnostic tests using receiver operating characteristics (ROC) analysis. *International Journal of Methods in Psychiatric Research, 2,* 39–50.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2006). *Measurement and assessment in education.* Boston, MA: Pearson.

Rumberger, R. W., & Larsen, K. A. (1998). Student mobility and increased risk of high school dropout. *American Journal of Education, 107,* 1–35.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior, 29,* 615–620.

Siegler, R. S. (2009). Improving the numerical understanding of children from low-income families. *Child Development Perspectives, 3,* 118–129.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428–444.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81,* 209–219.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Swets, J., & Picket, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Tabachnick, B. G., & Fidell, L. S. (2001). *Computer-assisted research design and analysis.* Boston: Allyn & Bacon.

Teacher Education Research Center. (1998). *Investigations in number, data, and space.* Columbus, OH: Scott Foresman/Addison-Wesley.

Teaching Integrated Math and Science Curriculum. (2004). *Math trailblazers.* Chicago: Kendall/Hunt.

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.

Watkins, M. W. (2005). Diagnostic validity of Wechsler subtest scatter. *Learning Disabilities: A Contemporary Journal, 3,* 20–29.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed., pp. 210–229). New York: Wiley.

Weiner, I. B. (2003). Prediction and postdiction in clinical decision making. *Clinical Psychology: Science and Practice, 10,* 335–338.

Weinstein, S., Obuchowski, N. A., & Lieber, M. L. (2005). Clinical evaluation of diagnostic tests. *American Journal of Roentgenology, 184,* 14–19.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 27,* 749–750.

Yovanoff, P., & Squires, J. (2006). Determining cut-off scores on a developmental screening measure: Use of receiver operating characteristics and item response theory. *Journal of Early Intervention, 29,* 48–62.

Date Received: July 2, 2009
Date Accepted: February 11, 1010
Action Editor: Tanya Eckert ■

Nancy C. Jordan, EdD, is Professor of Education at the University of Delaware. Her research interests include the development of number sense in children and learning difficulties in mathematics.

Joseph Glutting, PhD, is Professor of Education at the University of Delaware. He specializes in applied multivariate data analysis and test construction. He also is interested in the development of multivariate profile classification procedures and methods for testing hypotheses pertaining to profile uniqueness and prevalence.

Chaitanya Ramineni, PhD, is Associate Research Scientist in the Automated Scoring and Natural Language Processing Group at Educational Testing Services, Princeton, New Jersey. Her current research focuses on developing automated scoring capabilities for constructed-response scoring.

Marley W. Watkins, PhD, is Training Director and Professor in the School Psychology Program at Arizona State University. His research interests include professional issues, the psychometrics of assessment and diagnosis, individual differences, and computer applications.