Brendefur, J., Johnson, E. S., Thiede, K. W., Smith, E. V., Strother, S., Severson, H. H., & Beaulieu, J. (2015). Developing a comprehensive mathematical assessment tool to improve mathematics intervention for at-risk students. *International Journal for Research in Learning Disabilities, 2*(2), 65-90.

Developing a Comprehensive Mathematical Assessment Tool to Improve Mathematics Intervention for At-Risk Students

Jonathan Brendefur, Evelyn S. Johnson, and Keith W. Thiede
Boise State University

Everett V. Smith
University of Illinois, Chicago

Sam Strother
Boise State University

Herbert H. Severson
Oregon Research Institute

John Beaulieu
Visual Health Information, Tacoma, WA

Abstract

Students who complete kindergarten with an inadequate knowledge of basic mathematics concepts and skills will continue to experience difficulties with mathematics throughout their elementary and secondary years and may be at increased risk for math disabilities. There is a critical need to identify students experiencing difficulties in mathematics in the early elementary grades and to provide immediate and targeted instruction to remediate these deficits. Most early math screening tools focus on only a single skill, resulting in an incomplete picture of student performance and limited predictive validity. To address this need, we are developing a multiple-gating system of math assessment, the Primary Math Assessment (PMA), that both screens and provides diagnostic information in six domains. In this study, we present the results of the development and validation of items across the domains that will comprise

the PMA. Multidimensional Rasch models were used to estimate theoretically plausible dimensionality structures. Parsimony fit indices supported the six-dimensional model as the most generalizability model for the PMA data and supports reporting of six separate scores.

A recent review of early math screeners reported that virtually all screeners for the primary grades rely on assessing aspects of number sense (Gersten et al., 2012). These screening tools are used to identify students in need of more intensive math intervention as well as to identify students at risk for math disability. Two important issues present potential constraints on the efficacy of number sense screeners to adequately serve these purposes.

First, adoption of the Common Core State Standards (CCSS) in Mathematics has led to an expanded math curriculum that includes a much broader set of math content and process standards, especially in the primary grades (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). The Common Core is a set of high-quality academic standards outlining what a student should know and be able to do at the end of each grade. In math, the standards draw on the most effective international models for mathematical practice and include not only number and operations, but also algebraic thinking, measurement and data analysis, and geometry standards. Many primary-grade teachers do not have the mathematics knowledge to accurately identify students' needs across these more complex math skills (Hill, Rowan, & Ball, 2005) and, therefore, rely on established screening instruments to inform their decisions. Because the validity of instructional decisions depends to an extent on the alignment between the screening measures and the content standards on which classroom instruction is based (Irvin, Park, Alonzo, & Tindal, 2012), screening instruments that only assess number sense will not inform teachers of a student's skills across the other dimensions.

Second, math disability (MD) has typically been described as a core deficit in processing numerical quantity (Butterworth, Varma, & Laurillard, 2011) and number sense (Piazza et al., 2010; Wilson & Dehaene, 2007). Consistent with this description, the development of tools to identify children at risk for math disabilities has concentrated primarily on aspects of number sense (Gersten et al., 2012). However, a recent review of cognitive theories and functional imaging studies suggest that a model of number sense deficits as the unitary source of math disability is an oversimplification (Ashkenazi, Black, Abrams, Hoeft, & Menon, 2013). Specifically, in addition to a deficit in number sense, MD has also been described as (a) a specific impairment in symbolic processing and visual-spatial reasoning (Rousselle & Noel, 2007); (b) a domain-general deficit in working memory (Geary, 2004; Swanson, Howard, & Saez, 2006); and (c) a hybrid of impairments representing and manipulating numerical magnitude on an internal number line and in working memory and attention (Ashkenazi et al., 2013). Additionally, a majority of children with MD have

comorbid math and reading disabilities that result in significant difficulties with word problems (von Aster & Shalev, 2007). Each of these subtypes of math disability may result in impaired performance across a variety of math constructs and applications that are critical to identify in order to develop and implement interventions that meet an individual student's area of need.

Taken together, an increased understanding of various subtypes of math disability as well as significant changes to the math curriculum in the early grades call for screening and diagnostic tools that are more comprehensive than those currently available.

Improving Early Math Screening: The Primary Mathematics Assessment

The Primary Mathematics Assessment (PMA) (Brendefur & Strother, 2010) is being developed to address the major limitations of current early math screeners. The PMA is an assessment system designed for use in grades K-2 to identify students at risk for poor math outcomes across six dimensions and to provide further diagnostic information to guide intervention decisions. The PMA is designed as a multiple-gating system, in which students are first screened using the Primary Mathematics Assessment-Screener (PMA-S). Students who are identified as at risk on the screener are further assessed using the PMA-Diagnostic (PMA-D), a diagnostic assessment that provides a more complete evaluation of student performance to support intervention planning.

Multiple-gating approaches for identifying deficits in mathematics offer a promising solution to the problems with the "direct route" model of screening, in which intervention decisions are made based on the results of a screening test (Johnson, Jenkins, Petscher, & Catts, 2009). The goal of multiple gating is to administer a series of sequential assessments, in order to quickly assess a large population and identify students who have a high probability of being at risk for poor math performance. More in-depth evaluations are then used to confirm initial screening results and to provide a comprehensive analysis of a student's needs, which can then be used to inform intervention efforts. Multiple-gating approaches have been successfully applied to behavioral screening (Walker, Small, Severson, Seeley, & Feil, 2014) but are less commonly applied to academics. This is unfortunate because use of multiple-gating systems has demonstrated a reduction in intervention resource consumption by as much as 58% compared to single-stage screening procedures (Loeber, Dishion, & Patterson, 1984).

The PMA is hypothesized to measure six dimensions of math – number sequencing, operations (number facts), contextual problems, relational thinking, measurement, and spatial reasoning – that align closely with the Common Core State Standards in math and have been found to be highly predictive of later math achievement. A more complete review of the research supporting their importance for successful math achievement is presented elsewhere (Brendefur, Thiede, & Strother, 2012). Below, we provide a definition and brief synopsis of the research supporting each of the dimensions as critical components of early mathematics achievement.

Number sense/sequencing. Number sense has been suggested as the most important area of mathematical learning in early childhood (Clements & Sarama, 2007). This domain includes subitizing small quantities without counting, counting items in a set and knowing the final count word tells how many, discriminating between small quantities, comparing numerical magnitudes, and transforming sets of five or less by adding or taking away items (Jordan, Glutting, & Ramineni, 2008). A key component of number sense is counting or sequencing (Baroody, 1987). Counting has been described as the bridge between innate number sense and more advanced arithmetic abilities (Butterworth, 2004; Desoete, Ceulemans, Roeyers, & Huylebroeck, 2009). Given its role as a bridge to more advanced mathematics, sequencing is considered a prerequisite for future mathematical strategies such as basic operations (Blöte Lieffering, & Ouwehand, 2006; LeFevre et al., 2006). Several researchers (Geary, 2010; Geary, Hoard, & Hamson, 1999; Jordan, Glutting, & Ramineni, 2010) have found that difficulty in counting and other number sense deficits in early childhood is strongly predictive of later math achievement and should, therefore, be included as a key dimension on early math screeners.

Number facts. Basic math operations include the ability to add, subtract, multiply, and divide single digit numbers to 10. Fluency with math operations is critical for math achievement throughout students' school careers. Students with or at risk for math disabilities often have difficulty with fact retrieval, accurate computations (Geary, 2004) and flexibility, or the ability to solve problems in a variety of ways (Beishuizen & Anghileri, 1998). Immature calculating strategies, problems retrieving facts (Geary, 2004), and executive deficits (Passolunghi & Siegel, 2004) can prevent students from developing fluency with number facts (Geary, 2004), and result in more severe math learning challenges throughout school.

Contextual problems. Accurately solving contextualized problems is a key factor in early mathematics achievement, and word problems are a significant part of elementary math curricula. Contextualized problems serve as a means of developing students' general problem-solving skills and can promote proficiency with whole-number arithmetic (Verschaffel, Greer, & DeCorte, 2007). As described by Jitendra et al. (2013), contextualized problem solving requires the ability to understand the underlying problem type and related problem-solving procedures for that class of problems (Hatano, 2003), strong metacognitive skills (Montague, 2007), and the ability to distinguish relevant information (related to mathematical structure) from irrelevant details (Van Dooren, de Bock, Vleugels, & Verschaffel, 2010). Students at risk for learning disabilities that impact both math and reading (e.g., comorbid MD + RD) typically have difficulty solving word problems (Ashkenazi et al., 2013).

Relational thinking. Relational thinking describes the thinking of students who use number and operation sense to reflect on mathematical expressions as objects rather than as arithmetic procedures to be carried out (Carpenter, Franke, & Levi, 2003). As such, relational thinking is a

precursor to the development of algebraic thinking. Sarama and Clements (2008) stressed the importance of recognition and analysis of patterns in the early years to bring order and facilitate generalizations in math. An example of relational thinking provided by Stephens (2006) suggests that a student who is thinking relationally is able to recognize the equivalence of 3(x+4) and 3x+12 by attending to their structures without the need to solve the problem. The ability to find and extend numerical patterns to develop relational thinking is heavily dependent on how students are taught and should be a significant component of children's learning of mathematics (Sarama & Clements, 2009). Briefly, relational thinking is developed through helping students understand the equal sign (Driscoll, 1999), recognizing that equality is preserved if equivalent transformations are made on both "sides" of an equation, and can be fostered by posing true/false and open number sentences (Carpenter et al., 2003).

Measurement. Measurement of length has a direct link to understanding fractions and decimals because measurements often do not use complete units (Cramer, Post, & del Mas, 2002; Lehrer, 2003; Watanabe, 2002). For example, a table can be 3-½ feet wide. Students must make sense of the part of the unit left over after the three complete units are counted. Through this process, students develop a model for the continuous nature of rational numbers, which supports learning about fractions and ratios in later grades (Lehrer, Jaslow, & Curtis, 2003; McClain, Cobb, Gravemeijer, & Estes, 1999). Measurement tasks also support stronger proportional reasoning, which in turn supports understanding of geometry, numeracy, and data analysis (National Research Council, 2001). The underlying principles of measurement are unit iteration, partitioning, comparative measurement, and the meaning of measurement. Unit iteration is the act of repeating a unit to measure an object's attributes. Partitioning is the act of breaking an object into equal-sized measuring units (Lehrer, 2003). Finally, comparative measurement is the process of using a known measurement from one part of an object to find an unknown measurement (Kamii & Clark, 1997).

Spatial reasoning. Spatial reasoning is strongly correlated with achievement in math (Battista, 1981; Clements & Sarama, 2007; Gustafsson & Undheim, 1996). Students who perform well on spatial tasks also perform well on tests of mathematical ability (Geary, Hoard, Bryd-Craven, Nugent, & Numtee, 2007; Holmes, Adams, & Hamilton, 2008; McLean & Hitch, 1999). Spatial reasoning involves (a) spatial visualization, or the ability to mentally manipulate, rotate, twist, or invert pictures or objects; (b) spatial orientation, or the ability to recognize an object even when its orientation changes; and (c) spatial relations, or the ability to recognize spatial patterns, understand spatial hierarchies, and imagine maps from verbal descriptions (Lee, 2005). Recent evidence indicates that spatial reasoning training can have transfer effects on mathematics achievement, particularly on missing term problems (e.g. 7 + __ = 15), which are important for developing algebraic understanding (Cheng & Mix, 2014).

As demonstrated through this review of math constructs, a multidimensional measure of early math ability that can be efficiently administered and interpreted by elementary teachers

would allow more children with math deficits across a variety of important areas to be identified for intervention before these deficits begin to negatively affect math achievement. Additionally, a comprehensive measure such as the PMA would assist with intervention planning for children with specific deficits in critical math dimensions by providing a better match of intervention strategy to the demonstrated need.

Purpose of the Study

To develop a screening and diagnostic tool that adequately addresses the issues with existing math screeners, several phases of research have to be conducted. First, the psychometric qualities of the data, including estimates of reliability and investigations of dimensionality, need to be established. Next, the predictive validity of the screening tool needs to be established so that decision rules about performance can be tied to meaningful outcomes. Finally, the treatment validity of the assessment needs to be determined; that is, the extent to which the assessment contributes to positive outcomes (Gersten, Keating, & Irvin, 1995). Thus, there must be a clear and unambiguous relationship between the assessment data collected and the intervention that is recommended.

The current study reports on the development and validation of items for the PMA-S and PMA-D. The specific aims of the research were to:

1. Develop and determine the best set of items for assessing student ability within each of the six dimensions.
2. Assess the dimensionality of the PMA.
3. Determine the reliability of each of the six subscales.

Method

Participants

Students in kindergarten through second grade from seven schools within three districts in the Mountain Northwest participated in this project. All schools qualify for schoolwide Title 1 programs, with 40% or more of the student population eligible for free or reduced-price lunch; the number of students in kindergarten through sixth grade ranged from 350 to 450. All students were invited to participate, and those who returned parent consent forms participated. Across schools, between 70-74% of eligible students participated.

To prevent over-testing, not all participants responded to all of the questions. Classrooms were randomly assigned to complete two of the six dimensions. Three of the schools completed the number facts dimension in addition to their assigned dimensions. This

resulted in an uneven number of students completing the items to various dimensions. The demographics for the sample of students who participated in the data collection are presented in Table 1 by dimension. Students receiving special education services (i.e., students with individualized education program plans [IEPs]) were served under the category of speech or language impaired.

Table 1

Demographics by Subscale

	Number Sequencing	Number Facts	Contextual Problems	Relational Thinking	Measurement	Spatial Reasoning
Students (total)	97	232	124	112	119	131
Grade Level						
Kindergarten	28	61	37	31	36	39
First Grade	35	106	37	45	41	53
Second Grade	34	65	37	36	42	39
Sex						
Male	42%	39%	47%	46%	47%	55%
Female	58%	61%	53%	54%	53%	45%
Ethnicity						
Unspecified	14%	34%	27%	8%	29%	20%
American Indian	1%	0%	0%	0%	3%	2%
Asian	0%	1%	0%	2%	2%	3%
Black	2%	0%	1%	2%	1%	1%
Native	0%	0%	0%	0%	0%	1%
White	66%	43%	56%	70%	49%	50%
Hispanic	16%	19%	14%	17%	16%	24%
Multiracial	0%	3%	2%	2%	2%	1%
ELL	0%	0%	0%	1%	0%	1%
IEP	0%	0%	1%	4%	0%	3%

Note. ELL = English language learners. IEP = students with individualized education program plans (i.e., special education).

Procedures

The development and testing of the PMA took place over a nine-month period. A total of 148 items were created and dispersed across the six dimensions as follows: (a) 23 number sequence items, (b) 34 number fact items, (c) 10 context items, (d) 25 relational thinking items, (e) 25 measurement items, and (f) 31 spatial reasoning items. Items were administered to between 97 to 232 students in grades K-2, depending on the dimension. Rasch analysis (described in more detail in the data analysis section) allowed us to determine whether items fit the model requirements. Items were distributed across the six dimensions, with some items linked across grade levels as outlined in Table 2. Linked items are common items administered to more than one grade level so the calibration process would be able to place all items and persons across grades on a common metric.

Table 2

PMA Items Linked by Grade Level

	Sequencing	Facts	Relational Thinking	Context	Measurement	Spatial Reasoning	Total
Items per test/ grade level	23	34	25	10*	25	31	148
Items linked K – 1st grade	9	6	22	10	19	29	
Items linked 1st – 2nd grade	1	4	21	10	24	29	
Items linked K – 1st-2nd	0	1	17	10	18	27	

Measures

The PMA is designed as a multiple-gating system for students in kindergarten through second grade. The item bank will be used to develop a PMA-Screener and a PMA-Diagnostic. Students whose performance on the PMA-Screener indicates they may be at risk for poor performance in one or more dimensions will be evaluated using the PMA-Diagnostic in order to get a fuller evaluation of their abilities. Currently, we have developed and validated a total of 148 items across each of the six dimensions.

Data Analysis

Rasch models were used to analyze and evaluate the data. The use of Rasch models allows items and persons to be arranged in order of difficulty and ability, respectively, along a common metric, which in turn enables direct comparisons both between and across individuals and items. The metric can also be maintained across time points, which is necessary for understanding

which students require intervention and for calculating meaningful change scores over time once intervention has been provided. In addition, the Rasch model provides fit indices that aid in identifying items that may not contribute to measurement of the underlying dimension or latent trait measured. Finally, Rasch analysis provides person reliability indices that are analogous to internal consistency coefficients (KR20 and alpha, see Smith [2001] for why person reliability is a more accurate estimate of internal consistency than traditional estimates).

To determine the best fitting items to include on the PMA, the Rasch model were used to identify items that did not contribute to a unidimensional construct. For each subscale, we used WINSTEPS v3.8 (Linacre, 2014) to fit the data to a dichotomous Rasch (1-Parameter Logistic) model. This allowed us to (a) evaluate whether the items measured the desired constructs using item fit statistics (misfitting items were then revised or eliminated), (b) establish the internal consistency reliability of the subscales, and (c) use item Wright maps, provided by WINSTEPS to analyze the distribution of item difficulty with respect to the distribution of children's ability and remove items that were too difficult or too easy (i.e., poor targeting between item difficulty and children's ability).

To investigate the overall dimensionality of the data, multidimensional Rasch models for dichotomous data (Smith & Smith, 2004) were used for all Rasch modeling. The program used to estimate parameters for the multidimensional models was Conquest (Wu, Adams, & Wilson, 1997). We hypothesized the six dimensions of the PMA were statistically distinguishable. However, we also tested whether the dimensions could be combined to form a theoretically supported two-dimensional model consisting of Measurement and Spatial items as Dimension 1 and Sequencing, Facts, Contextualized Problems, and Relational Thinking as Dimension 2. Both hypothesized structures as well as a unidimensional model were evaluated using Conquest, which is capable of fitting multidimensional extensions of most basic unidimensional Rasch models using the Multidimensional Random Coefficient Multinomial Logit (MRCML) Model (Wu et al., 1997). Specifically, all three models were estimated and compared for relative model fit. However, the statistically best fitting model (i.e., minimizing the -2 log likelihood) does not mean that the identified model will be the model that generalizes the best as the model could overfit the data. As such, fit indices that take into account model complexity (e.g., number of parameters and/or number of observations) were implemented. These indices have collectively been labeled parsimony fit indices. The four parsimony fit indices employed for the dimensionality assessment were Akaike's Information Criterion (AIC; Akaike, 1974), the sample corrected Akaike Information Criteria (AIC-C; Burnham & Anderson, 2002), the Bayesian Information Criterion (BIC; Kass & Wasserman, 1995), and the Consistent Akaike Information Criterion (CAIC; Bozdogan, 1987). Lower values for these fit indices indicate the best tradeoff between model fit and generalizability.

Results

For each subscale, we first identified and removed misfitting items and investigated targeting issues. Item fit indices are expressed as mean squares, which represent the average value of squared residuals for each item, calculated from the difference between Rasch-predicted item performance and actual item performance in the observed data (Bond & Fox, 2013). Thus, larger mean square values represent poorer item fit with the Rasch model. The unstandardized unweighted mean square fit (MNSQ outfit) values have an expected value of 1. Values less than 1 indicate possible item redundancy or model overfit, whereas values greater than 1 indicate unpredictability or model underfit. In standardized form (ZSTD outfit), the expected value is 0 and approximates a unit normal distribution. Items for which the MNSQ outfit statistic was <1.3 or >0.7 and for which the ZSTD outfit was <2.00 or >-2.00 were considered to be fitting satisfactorily (Bond & Fox, 2013). We had an uneven number of participants across the various dimensions; however, mean square statistics have been found to be relatively independent of sample size when using polytomous data (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008).

Once the final set of items was developed, Wright maps displaying the item difficulty and student ability for each dimension were created (see Figures 1-6). As a means of explanation, we interpret the Wright map for measurement (see Figure 5). It displays the student ability measures expressed in logits (short for "log odd units," which result from applications of Rasch models) in a histogram on the left and the item difficult parameters on the right of the scale. The M, S, and T on either side of the vertical axis represent the mean, one standard deviation, and two standard deviations, respectively. The higher the student ability measure, the more able the student; the higher the item measure, the more difficult it is to get the item correct. Figure 5 indicates the items provide good coverage (e.g., targeting) for this sample of students, which helps contribute to a relatively small standard error of measurement for the student ability parameters. When item difficulties did not cover the range of student abilities, items were reviewed. For example, if there were too many students at the top of the map (ceiling effects) and too few items targeted toward these higher ability students, more difficult items were constructed.

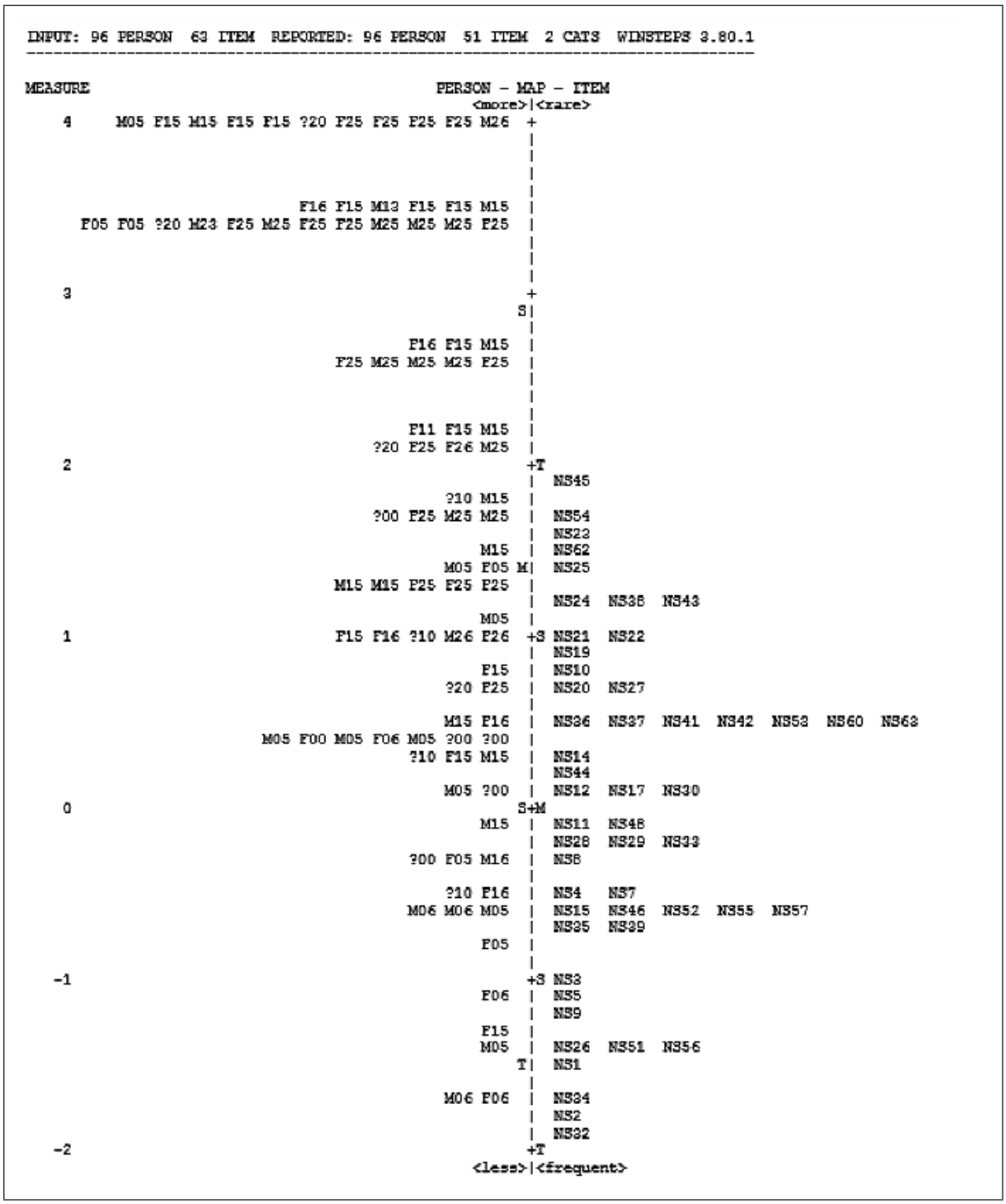

Figure 1. Wright map for number sequencing.

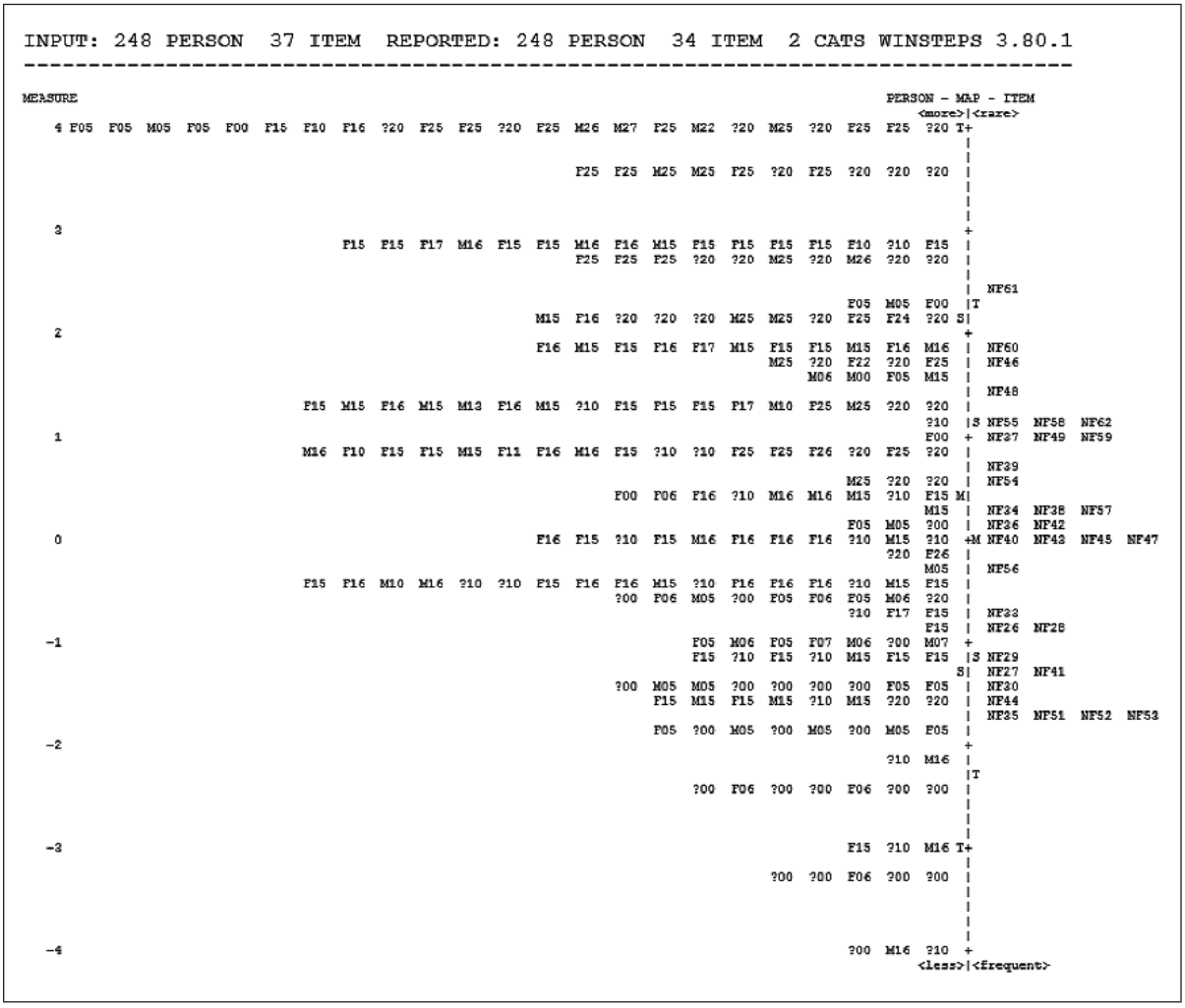

Figure 2. Wright map for number facts.

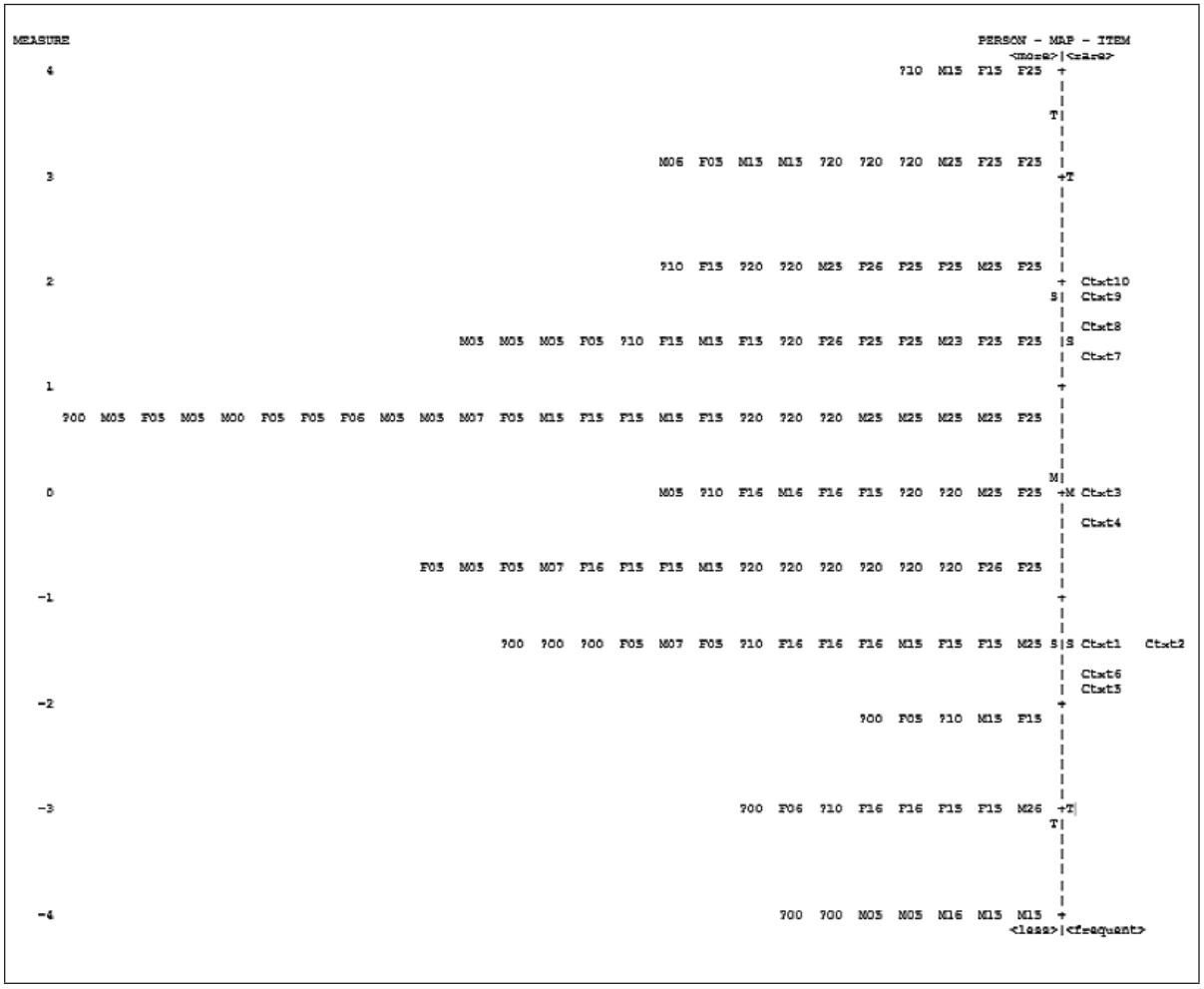

Figure 3. Wright map for contextualized problems.

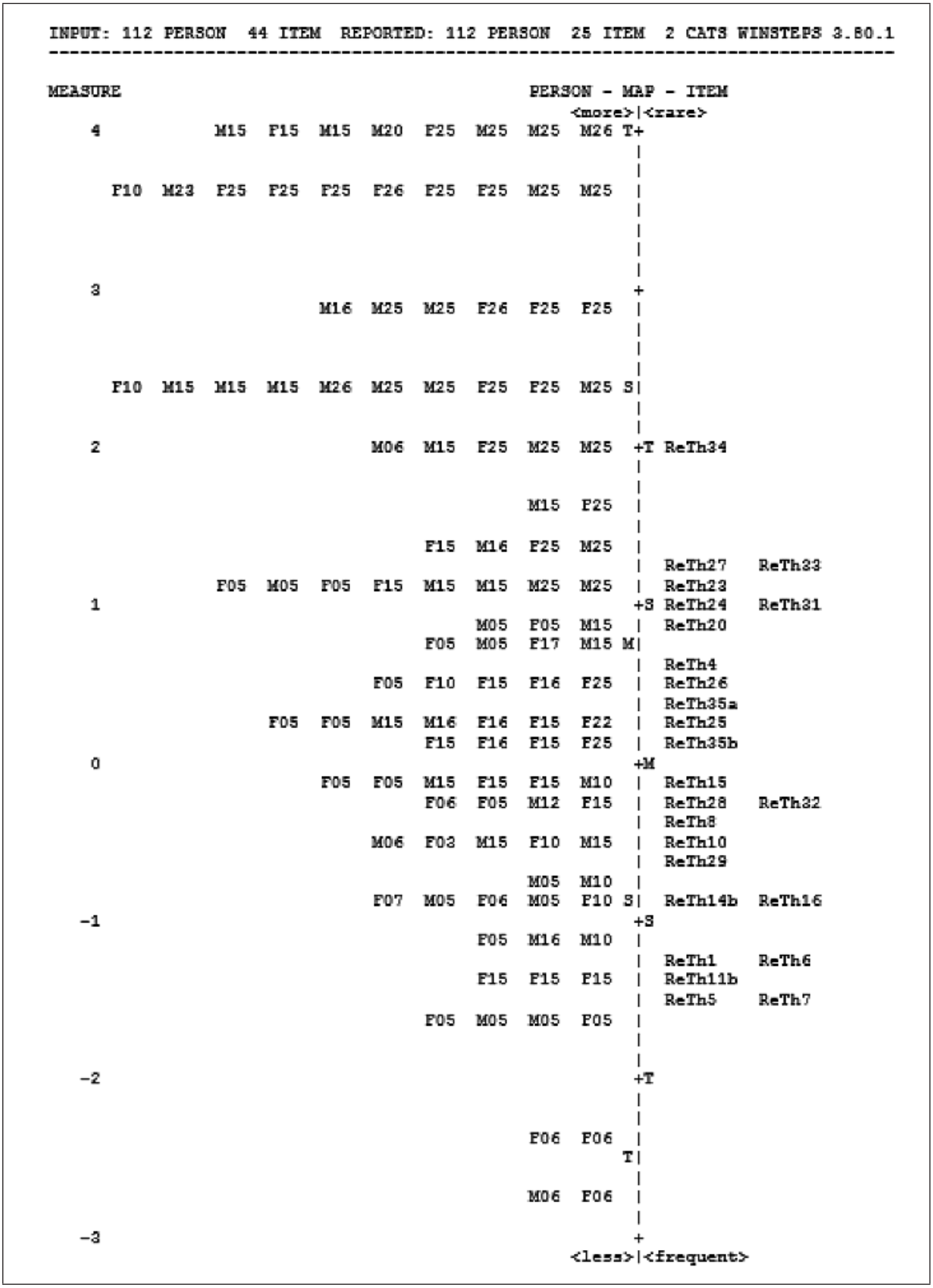

Figure 4. Wright map for relational thinking.

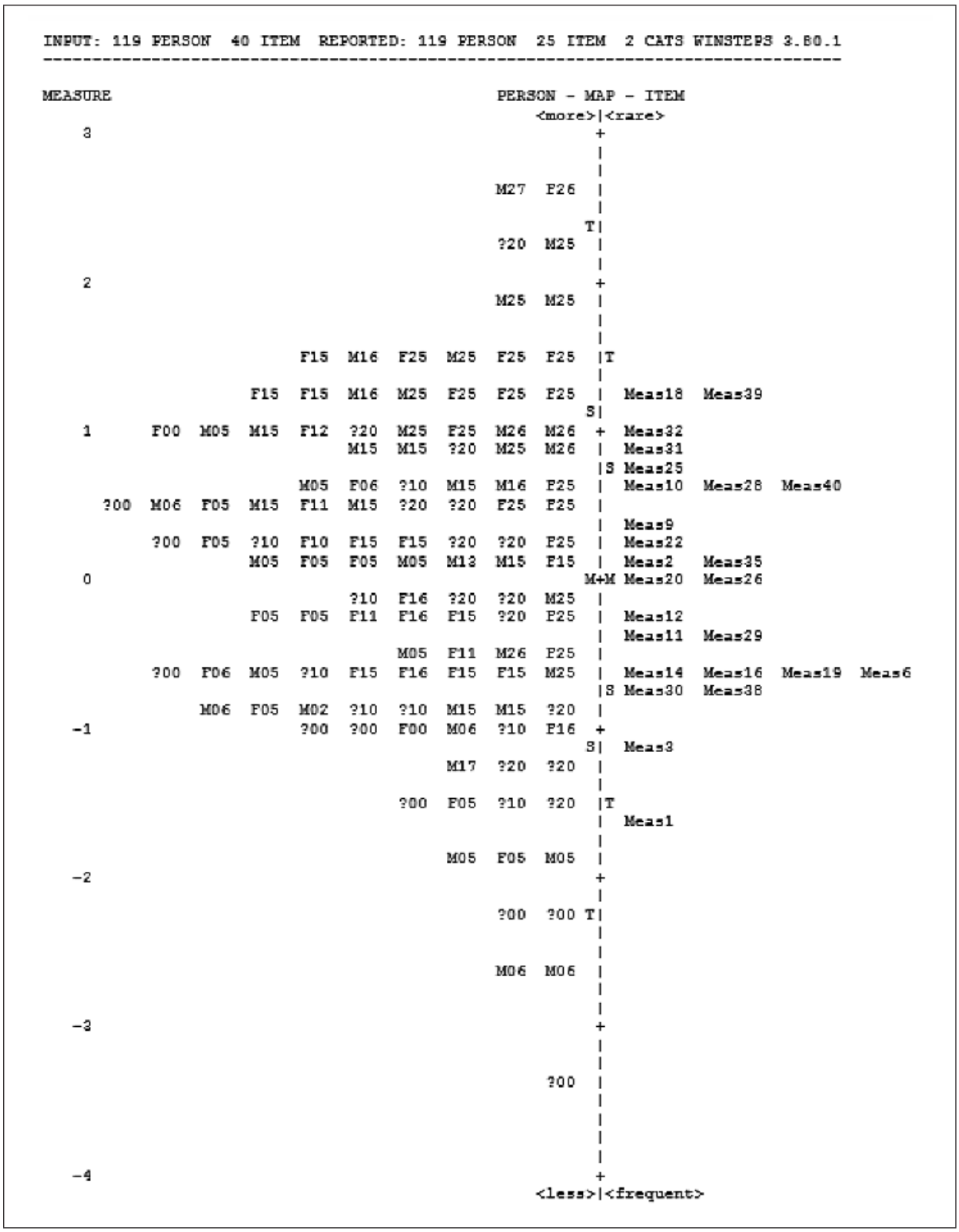

Figure 5. Wright map for measurement.

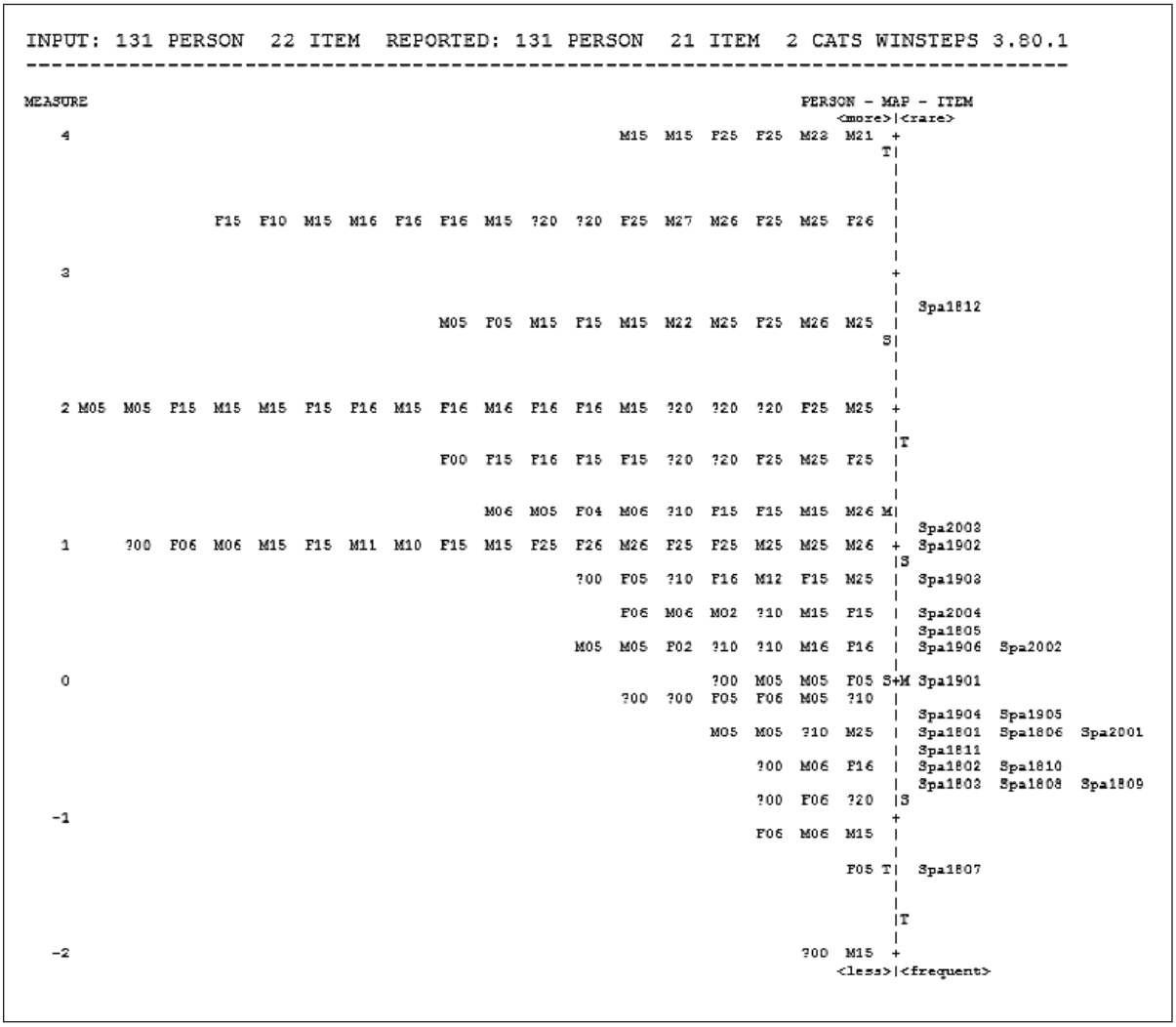

Figure 6. Wright map for spatial reasoning.

Rasch reliability. Rasch estimates of internal consistency reliability for items (see Table 3) and students (see Table 4) were also used to determine the quality of the data. For both persons and items, reliability of .70 to .79 is considered acceptable, .80 to .89 is good, and .90 or greater is excellent (Duncan, Bode, Lai, & Perera, 2003). As illustrated, in the current study, reliability for the various dimensions ranged from .74 to .87 for persons, and from .69 to .93 for items.

Table 3

Summary Statistics for Person Measures by Subscale on the Primary Math Assessment

Dimension		Infit		Outfit		Separation	Person Reliability
		MNSQ	ZSTD	MNSQ	ZSTD		
Number Sequencing (*n* = 97)	*Mean*	1.00	.0	1.06	.1	1.79	.78
	SD	.17	.7	.60	.9		
Number Facts (*n* = 232)	*Mean*	1.00	.0	1.02	.1	2.06	.81
	SD	.30	1.0	.69	1.0		
Context (*n* = 124)	*Mean*	.96	-.2	1.11	.5	1.77	.76
	SD	.13	1.0	.37	1.0		
Relational Thinking (*n* = 112)	*Mean*	1.00	.0	1.06	.0	2.53	.87
	SD	.20	1.0	.56	1.0		
Measurement (*n* = 119)	*Mean*	1.00	.0	1.00	.0	2.12	.82
	SD	.15	.8	.26	.9		
Spatial Reasoning (*n* = 131)	*Mean*	.99	.1	.99	.1	1.70	.74
	SD	.23	.7	.47	.8		

Note. MNSQ = unstandardized unweighted mean square fit values. ZSTD = standardized unweighted mean square fit values.

Table 4

Summary Statistics for Item Measures by Subscale on the Primary Math Assessment

Dimension		Infit		Outfit		Separation	Person Reliability
		MNSQ	ZSTD	MNSQ	ZSTD		
Number Sequencing (*n* = 23)	*Mean*	.97	-.1	1.02	.1	1.50	.69
	SD	.25	1.2	.58	1.1		
Number Facts (*n* = 34)	*Mean*	.00	.33	1.01	.0	2.93	.90
	SD	1.15	.14	.19	1.2		
Relational Thinking (*n* = 25)	*Mean*	.00	.25	-.2	1.06	3.65	.93
	SD	.98	.02	1.7	.63		
Measurement (*n* = 25)	*Mean*	.00	.21	.99	-.1	3.38	.92
	SD	.75	.01	.19	2.1		
Spatial Reasoning (*n* = 31)	*Mean*	.00	.23	1.00	.0	3.45	.92
	SD	.86	.02	.21	1.6		

MNSQ = unstandardized unweighted mean square fit values. ZSTD = standardized unweighted mean square fit values.

PMA construct structure evaluation. We hypothesized that a six-dimensional model would best fit the data. However, based on current conceptualizations of math disability as primarily related to number sense, we also hypothesized a two-dimensional model, in which sequencing, facts, context, and relational thinking would reflect one dimension around the construct of number sense and measurement and spatial reasoning would reflect a second dimension related to visual-spatial processing (R. Gersten, personal communication, June 2, 2014).

Table 5 lists the results of the multidimensional model comparisons. As illustrated, all four parsimony fit indices (i.e., AIC, AIC-C, BIC, and CAIC) favored the six-dimensional model, indicating that, among the models evaluated, the six-dimensional model is the most generalizable model for the PMA data. This finding was critical to supporting the conceptual framework of the PMA – the existence of six statistically distinguishable dimensions that can be used to inform a student's math ability and ultimately inform instruction and intervention needs across these distinct dimensions.

Table 5

Dimensionality of the PMA

Model	Deviance	Parameters	*N*	AIC	AIC-C	BIC	CAIC
6 dimensions	42940.54	235	751	43410.54	43625.92	44496.57	44731.57
2 dimensions	43518.43	217	751	43952.43	44129.93	44955.27	45172.27
Context	43999.14	215	751	44429.14	44602.75	45422.74	45637.74

Note. AIC = Akaike's Information Criterion; AIC-C = corrected Akaike Information Criteria; BIC = Bayesian Information Criterion; and CAIC = Consistent Akaike Information Criterion.

Discussion

As the evidence underscoring the importance of strong mathematics achievement continues to grow, more schools are realizing the need to begin instruction and intervention programs in the early grades to support students' growth and performance in math. Many early elementary teachers do not have strong foundations in teaching math, and lack the ability to accurately assess their students' instructional needs across a range of dimensions. This increases the likelihood that teachers use results from screening tools to inform their instruction. If the tools they use are one-dimensional (e.g., focus on number sense only), this may have the unintended consequence of restricting early mathematics instruction and intervention, leaving students unprepared for the demands of math instruction in later grades and possibly at increased risk of developing significant math learning challenges. Given the comprehensive nature of the

Common Core State Standards in Mathematics and our increased understanding of multiple subtypes of math disability, assessment tools that align with a broader set of mathematical constructs may provide an important way to help teachers ensure that their instruction is meeting the needs of their young students.

To address these concerns, we developed the Primary Math Assessment (PMA). The goal of the PMA is to create a multiple-gating assessment system to support the need for an efficient, accurate but comprehensive evaluation of K-2 students' math ability so appropriate instructional and intervention decisions can be made. The development of a multiple-gating assessment system requires several stages, including developing items and confirming that the dimensional model is consistent with the theoretical framework of the test, establishing the reliability and the validity of the screening and diagnostic results, and determining whether the use of the system has adequate treatment validity. The specific goals of this phase of PMA development were to (a) develop and determine the best set of items for assessing student ability within each of the six dimensions, (b) assess the dimensionality of the PMA, and (c) determine the reliability of each of the six subscales.

With regard to these goals, our analyses showed that the person reliabilities for relational thinking and measurement were greater than .80, and .81 for number facts, thus indicating that these dimensions can be reliably measured with the current items on the PMA. Other dimensions had person reliabilities ranging from .74 (spatial reasoning) to .78 (sequencing), suggesting that more items may be needed to reliably measure these dimensions. Our analysis of the dimensionality of the PMA indicated that the best fitting model for the PMA is a six-dimensional model, supporting the PMA's theoretical framework that a comprehensive assessment of multiple dimensions is important for informing student ability and subsequent intervention.

These findings are important in two ways. First, they provide evidence that there are ways to assess a broad set of skills that underlie critical math dimensions predictive of later math success. Second, as a first step in creating a system that will provide a quick, yet comprehensive assessment of student performance across a broad range of skills to inform instruction and intervention, the results reported here represent an encouraging improvement to currently available tools to address the needs of students at-risk for math disability. Being able to assess math performance in the very early elementary grades means that efforts to intervene will likely be more successful.

Limitations

Although the results reported in this manuscript are very encouraging, it is important to note that the data collection was conducted within one state and included seven schools from three districts. Although the demographics of the participating students includes high

percentages of students eligible for free or reduced-price lunch, English language learners (ELL) and a high percentage of Latino students, the demographics of the sample do not reflect those in other areas in the nation. As research and development of the PMA continues, a broader participant pool will be recruited.

Conclusion

It is evident that students in the early grades are not adequately prepared in mathematics (NCES, 2013). Using large data sets and nationally representative samples, several researchers have demonstrated that students who complete kindergarten with an inadequate knowledge of basic math concepts and skills will continue to experience difficulties with math throughout their elementary and secondary years (Duncan et al., 2007). This points to a critical need for early identification of students who are experiencing difficulties in math and subsequently to provide immediate and targeted intervention in order to build foundational skills and knowledge (Chernoff, Flanagan, McPhee, & Park, 2007). However, current screening tools tend to focus on few or single dimensions and are thus inadequate to fully inform early elementary teachers' instructional planning.

There is a great need and demand for reliable, efficient, and valid primary level math screening and diagnostic tools to identify students with math deficiencies so teachers can intervene with differentiated lessons in order to remediate student deficiencies. The results of this initial development indicate that assessment of the six dimensions included in the conceptual framework of the PMA may be reliably measured in the early grades to promote strong assessment and instructional planning to improve students' math proficiency. Next steps include developing and assessing new items and creating cut scores for the PMA-S and PMA-D that reliably identify students at risk for poor math achievement.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705, MR 0423716

Ashkenazi, S., Black, J. M., Abrams, D. A., Hoeft, F., & Menon, V. (2013). Neurobiological underpinnings of math and reading learning disabilities. *Journal of Learning Disabilities, 46*(6), 549-569.

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education, 18*(2), 141-157.

Battista, M. (1981). The interaction between two instructional treatments of algebraic structures and spatial-visualization ability. *The Journal of Educational Research, 74*(5), 337-341.

Beishuizen, M., & Anghileri, J. (1998). Which mental strategies in early number curriculum? A comparison of British ideas and Dutch views. *British Educational Research Journal, 24*(5), 519-538.

Blöte, A. W., Lieffering, L. M., & Ouwehand, K. (2006). The development of many-to-one counting in 4-year-old children. *Cognitive Development, 21*, 332-348.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

Brendefur, J. L., & Strother, S. (2010). *Idaho's Primary Mathematics Assessment: 2010 research report.* Boise, ID: Developing Mathematical Thinking, Inc.

Brendefur J. L., Thiede K., & Strother S. (2012). *Idaho's Primary Mathematics Assessment (PMA): Technical manual.* Boise, ID: Developing Mathematical Thinking Institute

Burnham, K. P., Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

Butterworth, B. (2004). The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry, 46*, 3-18.

Butterworth, B., Varma, S., & Laurillard, D. (2011). Dyscalculia: From brain to education. *Science, 332*(6033), 1049-1053. doi:332/6033/1049[pii]10.1126/science.1201536

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in the elementary school.* Portsmouth, NH: Heinemann.

Cheng, Y. L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development, 15*(1), 2-11.

Chernoff, J. J., Flanagan, K. D., McPhee, C., & Park, J. (2007). *Preschool: First findings from the third follow-up of the Early Childhood Longitudinal Study, Birth Cohort* (ECLS-B) (NCES 2008-025) . Washington, DC: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.

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.

Cramer, K. A., Post, T. R., & del Mas, R. C. (2002). Initial fraction learning by fourth- and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the Rational Number Project Curriculum. *Journal for Research in Mathematics Education, 33*(2), 111-144.

Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning. *Educational Research Review, 4*(1), 55-66.

Driscoll, M. (1999). *Fostering algebraic thinking: A guide for teachers, grades 6-10*. Portsmouth, NH: Heinemann.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Japeli, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428-1446.

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950-963.

Geary, D. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4-15.

Geary, D. C. (2010). Mathematical disabilities: Reflections on cognitive, neuropsychological, and genetic components. *Learning and Individual Differences, 20,* 130-133.

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.

Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: Patterns of functions and deficits in children at risk for a mathematical disability. *Journal of Experimental Child Psychology, 74,* 213-239.

Gersten, R., Keating, T., & Irvin, L. K. (1995). The burden of proof: validity as improvement of instructional practice. *Exceptional Children, 61*(6), 510-519.

Gersten, R., Clarke, B., Jordan, N., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children, 78*(4), 423-445.

Gustafsson, J.-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. Handbook of educational psychology. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186-242). London, England: Prentice Hall International.

Hatano, G. (2003). Foreword. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills* (pp. xi-xiii). Mahwah, NJ: Erlbaum.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effect of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Holmes, J., Adams, J. W., & Hamilton, C. J. (2008). The relationship between visuospatial sketchpad capacity and children's mathematical skills. *European Journal of Cognitive Psychology, 20*, 272-289.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). *The alignment of the easy CBM grades 6-8 math measures to the Common Core Standards* (Report No. 1230). Eugene, OR: University of Oregon, Behavioral Research and Teaching.

Jitendra, A. K., Dupuis, D. N., Rodriguez, M. C., Zaslofsky, A. F., Slater, S., Cozine-Corroy, K., & Church, C. (2013). A randomized controlled trial of the impact of schema-based instruction on mathematical outcomes for third-grade students with mathematics difficulties. *The Elementary School Journal, 114*(2), 252-276.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). Screening for early identification and intervention: How accurate are existing tools and procedures in predicting first grade reading outcomes? *Learning Disabilities Research and Practice, 24*(4), 174-194.

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.

Kamii, C., & Clark, F. (1997). Measurement of length: The need for a better approach to teaching. *School Science and Mathematics, 97*, 116-121.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of American Statistical Association, 90*, 928-934.

Lee, J. W. (2005). *Effect of GIS learning on spatial ability* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.

LeFevre, J. A., Smith-Chant, B. L., Fast, L., Skwarchuk, S. L., Sargla, E., Arnup, J. S. et al. (2006). What counts as knowing? The development of conceptual and procedural knowledge of counting from kindergarten through grade 2. *Journal of Experimental Child Psychology, 93*, 285-303.

Lehrer, R. (2003). Developing understanding of measurement. In J. Kilpatrick, M. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 179-192). Reston, VA: National Council of Teachers of Mathematics.

Lehrer, R., Jaslow, L., & Curtis, C. (2003). Developing an understanding of measurement in the elementary grades. In D. H. Clements & G. Bright (Eds.), *Learning and teaching measurement* (pp. 100-121). Reston, VA: National Council of Teachers of Mathematics.

Linacre, J. M. (2014). *Winsteps*® (Version 3.81.0) [Computer software]. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com/

Loeber, R., Dishion, T. J., & Patterson, G. R. (1984). Multiple gating: A multistage assessment procedure for identifying youths at risk for delinquency. *Journal of Research in Crime and Delinquency, 21*, 7-32.

McClain, K., Cobb, P., Gravemeijer, K., & Estes, B. (1999). Developing Mathematical Reasoning Within a Context of Measurement. In L. Stiff (Ed.), *Developing mathematical reasoning, K-12* (pp. 93-106). Reston, VA: National Council of Teachers of Mathematics.

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology, 74*, 240-260.

Montague, M. (2007). Self-regulation and mathematics instruction. *Learning Disabilities Research & Practice, 22*(1), 75-83.

Primary Math Assessment by Jonathan Brendefur, Evelyn S. Johnson, Keith W. Thiede, Everett V. Smith, Sam Strother, Herbert H. Severson, and John Beaulieu

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics.* Washington, DC: Authors.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Passolunghi, M. C., & Siegel, L. S. (2004). Working memory and access to numerical information in children with disability in mathematics. *Journal of Experimental Child Psychology, 88*, 348-367.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., Dehaene, S., & Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition, 116*, 33-41.

Rousselle, L., & Noel, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic versus non-symbolic number magnitude processing. *Cognition, 102*(3), 361-395. doi:10.1016/j.cognition.2006.01.005

Sarama, J., & Clements, D. (2009). *Learning and teaching early math: The learning trajectories approach*. New York, NY: Routledge.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(33). doi:10.1186/1471-2288-8-33

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and the validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281-331.

Smith, E. V., & Smith, R. M. (2004). *Introduction to Rasch measurement.* Maple Grove, MN: JAM Press.

Stephens, A. C. (2006). Equivalence and relational thinking: Preservice elementary teachers' awareness of opportunities and misconceptions. *Journal of Mathematics Teacher Education, 9*, 249-278.doi:10.1007/s10857-006-9000-1

Swanson, H. L., Howard, C. B., & Saez, L. (2006). Do different components of working memory underlie different subgroups of learning disabilities? *Journal of Learning Disabilities, 39*(3), 252-269.

Van Dooren, W., de Bock, D., Vleugels, K., &Verschaffel, L. (2010). Just answering . . . or thinking? Contrasting pupils' solutions and classifications of missing-value word problems. *Mathematical Thinking and Learning, 12*, 20-35. doi:10.1080/10986060903465806

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operation. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 557-628). Charlotte, NC: Information Age.

Von Aster, M., & Shalev, R. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology, 49*(11), 868-873. doi:DMCN868 [pii] 10.1111/j.1469-8749.2007.00868.x

Walker, H. M., Small, J. W., Severson, H. H., Seeley, J. R., & Feil, E. G. (2014). Multiple-gating approaches in universal screening within school and community settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 47-75). Washington, DC: American Psychological Association.

Watanabe, T. (2002). Representations in teaching and learning fractions. *Teaching Children Mathematics, 8*(8), 457-563.

Wilson, A. J., & Dehaene, S. (2007). Number sense and developmental dyscalculia. In D. Coch, K. Fischer, & B. Dawson (Eds.), *Human behavior and the developing brain: Atypical development* (2nd ed., pp. 212-238). New York, NY: Guilford Press.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wu, M., Adams, R. J., & Wilson, M. R. (1998*). ACER ConQuest: Generalized item response modeling software* [Computer program and manual]. Camberwell, VIC, Australia: Australian Council for Educational Research.