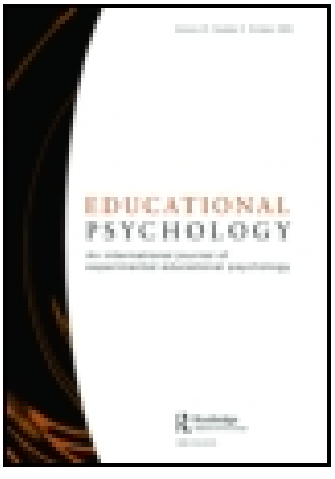


Educational Psychology: An International Journal of Experimental Educational Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cedp20

Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment

Douglas H. Clements [a], Julie H. Sarama [a] & Xiufeng H. Liu [a]

[a] Learning and Instruction, School of Education, University of Buffalo, Buffalo, USA

Published online: 20 May 2008.

To cite this article: Douglas H. Clements, Julie H. Sarama & Xiufeng H. Liu (2008) Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment, Educational Psychology: An International Journal of Experimental Educational Psychology, 28:4, 457-482, DOI: 10.1080/01443410701777272

To link to this article: http://dx.doi.org/10.1080/01443410701777272

Educational Psychology
Vol. 28, No. 4, July 2008, 457–482

Routledge
Taylor & Francis Group

Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment

Douglas H. Clements*, Julie H. Sarama and Xiufeng H. Liu

Learning and Instruction, School of Education, University of Buffalo, Buffalo, USA

(*Received 16 June 2007; final version received 29 October 2007*)

There are only a few instruments to assess mathematics knowledge and skills in children as young as three to four years of age, and these instruments are limited in scope of content. We describe the development of a theoretically based, empirically tested instrument designed to measure the mathematical knowledge and skills of children from three to seven years of age, emphasising its submission to the Rasch model. After using the data to refine the instrument, they fit the model well, with high reliability. These data also provided empirical support for the developmental progressions for most topics. We conclude with a description of the research's contribution to theory and empirical research regarding young children's development of specific mathematical competencies.

Keywords: academic performance; test; developmental; maths

There has been a surge of interest and activity in the area of early childhood mathematics. However, there is a paucity of assessment instruments appropriate for children as young as three to four years of age, and the few that have been created are limited either in content scope or in their development *qua* assessment instruments. Here we describe the development of a theoretically based, empirically tested instrument designed to measure the mathematical knowledge and skills of children from three to seven years of age, emphasising its submission to the Rasch model and its contribution to theory and empirical research regarding young children's development of specific mathematical competencies.

Background

There are at least four reasons for the burgeoning interest in early mathematics. First, increasing numbers of children attend early care and education programmes (Barnett, Hustedt, Hawkinson, & Robin, 2006; Doig, McCrae, & Rowe, 2003; Hinkle, 2000). Second, there is an increased recognition of the importance of mathematics (Doig et al., 2003; Kilpatrick, Swafford, & Findell, 2001) and the varying performance in mathematics of students in different countries (Mullis et al., 1997), beginning in the preschool years (Blevins-Knabe & Musun-Miller, 1996; Ginsburg & Russell, 1981; Griffin, Case, & Capodilupo, 1995; Holloway, Rambaud, Fuller, & Eggers-Pierola, 1995; Jordan, Huttenlocher, & Levine, 1992; Saxe, Guberman, & Gearhart, 1987; Starkey, Klein, Chang, Lijuan & Yang, 1999). Third, there are differences not just between nations, but also between low- and higher-income children (Thomson, Rowe, Underwood, & Peck, 2005) and different ethnic and cultural groups (National Center for Education Statistics, 2000; Thomson et al., 2005). Fourth, government agencies provide financial support for pre-kindergarten programmes designed to

*Corresponding author. Email: clements@buffalo.edu

ISSN 0144-3410 print/ISSN 1469-5820 online

DOI: 10.1080/01443410701777272
http://www.informaworld.com

facilitate academic achievement, particularly for low-income children, and research indicates that early interventions in mathematics can prevent later learning difficulties in school for all children (Clements & Sarama, 2007c; Doig et al., 2003; Fuson, Smith, & Lo Cicero, 1997; Griffin, 2004; Wright, 2003).

Such attention to early mathematics generates a need for assessment instruments. Extending our knowledge of young children's mathematical development, and evaluating the effectiveness of programmes designed for them, requires accurate measures of mathematical knowledge and skill. The available instruments are useful for certain purposes, but are limited in several ways – limitations we have attempted to surmount in designing and formatively evaluating the instrument described here.

For example, a common instrument is the Woodcock–Johnson III (Woodcock, McGrew, & Mather, 2001), and in particular its 'applied problems' section. This subtest has several strengths: it assesses a wide range of abilities and ages, has reliabilities above .80, and there are large normative data samples. However, two national panels on preschool assessment (NICHD Forum, Washington, DC, June 2002; CIRCL Forum, Temple University, January 2003) cautioned against sole use of the Woodcock–Johnson for the assessment of mathematical skills in preschoolers for the following reasons: it has not been validated for children in the youngest age ranges; it covers a narrow range of problems (e.g., the oft-used 'applied problems' subtest has multiple tasks in which children must count a proper subset, all with numbers from 1 to 4) and jumps too quickly to advanced, formal knowledge; and it is not based on current research into the development of mathematical thinking – for example, there is little attention to developmental sequences.

Other assessments have similar limitations. For example, the Bracken Basic Concept Scale (Bracken, 1984/1998) includes several mathematical concept categories; however, the national panels cautioned that content validity was low (including mathematically questionable items such as one-dimensional 'linear shapes'), and, because subtests were not intended to be used separately, it can be difficult to administer or interpret results for mathematical topics.

A measure more positively reviewed by the national panels is the Test of Early Mathematics Ability (Ginsburg & Baroody, 2003), which has adequate internal consistency and validity. However, although the official description of the instrument states that it 'measures the mathematics performance of children', this is only partially true. It measures number competencies, but not the other important topics in mathematics. Another early mathematics assessment measures several underlying competencies and learned skills, and classifies children into five developmental levels, but is not designed to request the correct answers to specific questions (de Lemos & Doig, 1999). A combination of this and another measure (Doig & de Lemos, 2000) was analysed using the Rasch model, and yielded a reliable and valid Pre-school Numeracy Scale (Thomson et al., 2005), which lacks only fine-grained developmental progressions for various mathematical topics.

In summary, there is a need for a single, theoretically-based instrument that addresses the range of mathematical thinking of which young children are capable, including multiple mathematical topics and the developmental sequences within each. Our theoretical framework is hierarchic interactionalism, the first tenet of which is the notion of developmental progressions:

> Most content knowledge is acquired along developmental progressions of levels of thinking…. [These] play a special role in children's cognition and learning because they are particularly consistent with children's intuitive knowledge and patterns of thinking and learning at various levels of development. (Clements & Sarama, 2007b, p. 464, see this chapter for a full description of the theory and its tenets)

Previous research supports the premise that different categories of mathematical knowledge grow in parallel (Doig & de Lemos, 2000; Van de Rijt & Van Luit, 1999), which is the basis of topic-specific developmental sequences in our theoretical and empirical approach. As an example, one

such developmental progression involves shape composition. Born in observations of children's explorations (Sarama, Clements & Vukelic, 1996), this developmental progression was refined through a series of clinical interviews and focused observations (Clements, Sarama & Wilson, 2001), and validation studies (Clements, Wilson & Sarama, 2004). From a lack of competence in composing geometric shapes, children gain the ability to combine shapes – initially through trial and error, and gradually by attributes – into pictures, and finally synthesise combinations of shapes into new shapes (composite shapes). Other developmental progressions were based on extensive reviews of the research (Clements & Sarama, 2007b). These provided the theoretical and empirical core around which we designed, evaluated, and revised the measure of mathematical knowledge and skills of children from three to seven years of age. The remainder of this paper discusses the development of this instrument, especially a large-scale implementation that yielded information on the psychometrics of the instrument and on the developmental progressions it was designed to measure.

Method

Development and initial qualitative testing of the instrument

The Research-Based Early Maths Assessment (REMA) measures preschool children's mathematical knowledge and skills. We determined the content goals by considering what mathematical topics are valued by educators, mathematicians, and researchers in the field (e.g., National Council of Teachers of Mathematics, 2000, 2006; Van de Rijt, Van Luit, & Pennings, 1999), and the empirical research on what constitutes the core ideas and skill areas of mathematics for young children (e.g., Baroody, 2004; Clements, Sarama, & DiBiase, 2004; Fuson, 1997), evaluated by the criteria that they be 'mathematically central and coherent, consistent with children's thinking, and generative of future learning' (Clements et al., 2004, p. 13). These areas were: number, including verbal counting, object counting, number recognition and subitising, number comparison, number sequencing, numeral recognition, number composition and decomposition, and adding and subtracting; geometry, including shape identification, shape composition and decomposition, comparison and congruence, construction of shapes, and transformations; measurement; and patterns. Note that general concepts and processes, such as part–whole thinking, and the corresponding processes of composition and decomposition, classification, and seriation were woven throughout several areas. Classification was taken as the appropriate foundational skill for the domain of data analysis; no other concepts or skills from such areas as data analyses, probability, or fractions passed the criteria for selection for this age level.

For each mathematical topic, we reviewed the research (see reviews in Clements & Sarama, 2007b; Clements, Sarama & DiBiase, 2004) to determine a developmental progression. We elaborated each developmental progression as necessary to include specific behavioural indicators for each level of thinking within each topic (for a complete list, see Clements & Sarama, 2007a, pp. B2–B14). We generated items by creating or adapting tasks designed to elicit such behaviours, with no fewer than two for each level. The sequence of items was determined by their position in the developmental progression.

The initial assessment was refined in three pilot tests. The first two tests were conducted to evaluate individual items formatively, in terms of both their ease of administration and the extent to which qualitative observations validated each item's ability to assess a given level of thinking (Clements & Sarama, 2004; Sarama & Clements, 2002). We then arranged a review by an expert panel to ensure content validity: the panel members were Les Steffe (mathematics and mathematics education), Mary Lindquist (mathematics and early childhood education), Rene Parmar (assessment, special education, cognitive psychology), and Chuck Thompson (early childhood and mathematics education). Their reports were generally positive, and each criticism or suggestion was

followed, including the presentation of a scenario for the number items (shopping for groceries) and the addition of several items.

In the third pilot test phase, we administered the revised instrument to another cohort of three to five year olds (Clements & Sarama, 2007c). All assessments were videotaped and subsequently coded by independent coders, who had been previously trained and evaluated. Codes included correct/incorrect answers and separate codes for children's strategies in cases where those strategies were intrinsically related to the level of thinking the item was designed to measure. Coefficient alpha reliabilities ranged from $r = .89$ (number) to .71 (geometry); interrater reliability was 98%. Concurrent validity of the total test score was established with a .86 correlation with another measure of preschool mathematics achievement (Klein, Starkey, & Wakeley, 2000).

Although this version of the test was acceptable, there were several limitations. First, except for the longest developmental progressions (object counting and counting strategies; arithmetic), the stop rule (the assessment for each progression was stopped when a child gave three consecutive incorrect answers) was rarely evoked, as there were too few items in any one topic. Thus, the sections took considerable time to administer – often a total of 60 minutes or more for each child. Second, total scores based on a simple sum of the item scores would be based on weak theoretical ground. There was no reason to suspect that the instrument yielded an equal interval scale. This is important for any instrument that is to be used to evaluate a curriculum or programme.

For these reasons, we revised and evaluated the instrument. We defined mathematical competence as a latent trait using the Rasch model (Bond & Fox, 2001; Linacre, 2005; Watson, Callingham, & Kelly, 2007), leading to scores that locate children on a common ability scale with a consistent, justifiable metric, allowing accurate comparisons, even across ages, and meaningful comparison of change scores, even when initial (e.g., pretreatment) scores differ (Wright & Stone, 1979). We re-sequenced the items, strictly maintaining the order within each topic, but intermingling items across topics according to the available developmental evidence – both age specifications from the literature and difficulty indices from our pilot testing. Thus, item were posited to be in an increasing order of difficulty, but theoretical claims that this sequencing represented increasingly sophisticated levels of mathematical thinking were made only for items within a given topic. We submitted the results from administering this revised instrument to the Rasch model.

Participants

The population came from diverse classrooms serving preschoolers in New York state. The first group included 34 low-income classrooms with 99% (Head Start, a programme of the US Department of Health and Human Services that provides education for children from low-income families) and 74% (state-funded, in which individual states set criteria and fund programmes that may be housed in public schools or other settings) children receiving reduced or free lunch and 70% (Head Start) and 72% (state-funded) minority children (58%, 11%, and 3%, and 47%, 13%, and 10%, African-American, Hispanic, and Other, for the two programmes, respectively). Teachers in these schools were 36% (Head Start) and 19% (state-funded) minority, with a median of 8 (Head Start) to 16 (state-funded) years' experience; 28% (Head Start) to 90% (state-funded) have NYS certification. The second group included 12 mixed-income classrooms, averaging 19% free or reduced lunch, and 30% minority (records allowed no additional categorical breakdowns; they did reveal that no children on reduced lunch were on public assistance). In each classroom, we randomly selected about eight children from the pool of all kindergarten-intending (for kindergarten in 2004–2005) preschoolers who returned Institutional Review Board permission forms. A few Head Start classrooms had only eight, or sometimes fewer, kindergarten-intending children; in those cases, all those qualifying were tested. Four children moved out of the area during

the study, resulting in a total of 360 children. Each was assessed twice; at the beginning of the school year (60 were age three and 300 were age four) and at the end of the school year (211 were age four and 149 were age five).

Procedure and analyses

Doctoral students in educational psychology were trained on the REMA. They practiced on children not included in the study, until they demonstrated no experimenter errors on three consecutive administrations. They then assessed children at the beginning and end of their preschool year.

To capture performance on as many items as possible, without stressing the children, the stop rule was set at double the number of items as had been used in previous administrations; that is, the assessment stopped only when a child was completely incorrect on six consecutive items. Child mathematics outcomes were coded and then scored by trained teams (not the assessors) naïve to the children's treatment group.

Rasch analyses can be used to estimate the distance between items, as well as between persons, on a single scale. Rasch developed a probabilistic model in which item difficulty (a test item's underlying difficulty based on the proportion of a given sample that responded correctly) and person measure (a person's underlying competence, based on the proportion of items completed correctly) are simultaneously estimated. The result is a scale on which both persons and items are mapped onto the theoretical latent trait (an attribute of persons, in our case mathematical competence, that can be inferred from behaviour) in the same equal-interval units. These are log-odds probability units, or logits. Because logits are inconvenient to use (e.g., with negative values), Wright and Stone (1979) recommend rescaling logits into a more meaningful scale. In this study, we chose to rescale the logits to scale-equivalent *t* scores, with $M = 50$, $SD = 10$ for ease of interpretation (Linacre, 2005). Several items on our assessment had scoring protocols that specified multiple levels of credit (e.g., verbal counting was initially scored into five levels: 0; correct counting to 5 = 1 point; correct counting to 10 = 2 points; correct counting to 20 = 3 points; correct counting to 35 = 4 points). Therefore, we submitted the scores to a partial credit Rasch model, which takes the following form:

$$\ln\left(\frac{P_{nik}}{1-P_{nik}}\right) = B_n - D_i - F_{ik}$$

where P_{nik} is the probability of person n responding at level k to item i. Thus, the left-hand side of the equation represents a person's likelihood or log-odds for answering a question at a particular level (e.g., right, partially right, or wrong). P_{nik} ranges from 0 to 1. B_n is person n's latent ability; it is usually expressed in logits. Thus, B_n ranges theoretically from $-\infty$ to $+\infty$ but usually from −5 to +5 in a given group due to the limited variation within the group. D_i is the average difficulty of item i, and F_{ik} is the difficulty of level k for item i. Difficulty level represents the threshold at which a person with an ability estimate equivalent to the item difficulty has a 50% probability of responding correctly. Placed on the same scale as ability (B_n), item difficulty (D_i and F_{ik}) ranges from $-\infty$ to $+\infty$ with more negative values indicating easier items and more positive values indicating more difficult items. Because B_n and D_i/F_{ik} are on the same scale, student latent abilities are directly comparable to item difficulties.

We used the Winsteps computer programme (Linacre, 2005) to estimate item difficulties, fit statistics, reliabilities, and separation indices. Fit statistics are estimates of the degree to which responses show adherence to the expectations of the Rasch model. They indicate how well the model's assumption of unidimensionality is empirically supported – that is, whether the instrument is measuring a single attribute (a critical characteristic of true measurement). Item reliability

is an estimate of the replicability of item placement within the hierarchy of items along the measured trait, and is conceptually similar to Cronbach's alpha. Item separation indices are estimates of the spread of items, or the ability of the measure to differentiate items.

Because qualitative analyses had been employed to refine the instrument in several previous cycles of formative testing and revision, qualitative examination of video tapes was reserved for items with poor item characteristics.

Of the original 236 items, 216 were scored dichotomously. The rest scored in three to five categories, and were assigned a weight of 2. All dichotomously scored items, except items assessing shape recognition, were weighted equally (i.e., a weight of 1). The geometry tasks involved setting out 26 shapes (starting with all shapes each time), and asking the children to select all the squares, triangles, rectangles, or rhombuses. Each shape was an item, weighted as follows (with examples for squares): palpable distractors (those without any overall resemblance such as ovals), .05; exemplars (theoretically- and empirically-determined prototypes of the class, such as square with a horizontal base), .25; 'variants' (other members of the class, such as a square with sides at 45°), .50; and distractors (non-examples that are easily confused with examples, such as a rhombus with angles of 60° and 30°), also .50. The total of these weightings was approximately 3 for each of the four shape recognition tasks.

Results

The purpose of the study was to improve an instrument that measures the mathematical knowledge and skills of young children. We used the Rasch model to refine, then validate, our assessment approach.

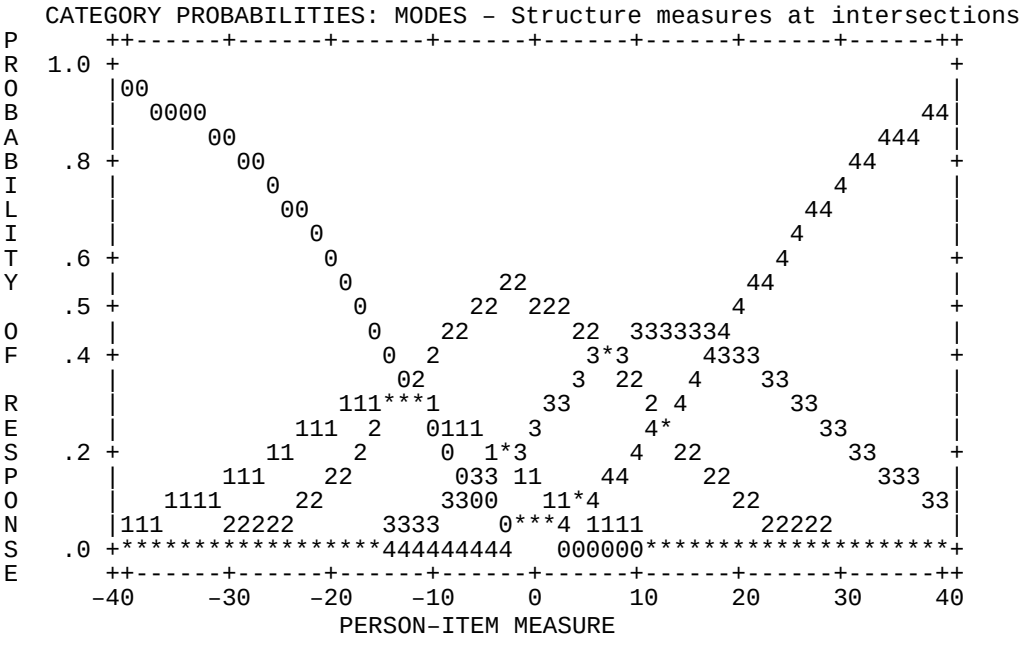
Evaluation and revision of individual items

We revised the coding of several items, due to their poor category structure as revealed through item characteristic curves, plots of the probabilities of a correct response for each value of the person measure. For example, number item 1, which simply asked children to count verbally without objects, was originally coded as 0 for no counting, 1 for correct counting to 5, 2 for correct counting to 10, 3 for correct counting to 20, and 4 for correct counting to 35 or beyond. Figure 1a shows the original category structure. Although the item characteristic curves for codes 0, 2, and 4 are acceptable, code 3 does not discriminate well, and code 1 is not useful in representing a distinct response category. We recoded the item, combining the old codes 1 and 2 (creating a new code 1) and the old codes 3 and 4 (creating a new code 2), which yielded a more satisfactory category structure, as illustrated in Figure 1b. Of 7 number and 11 other (geometry, measurement, patterns) items with non-dichotomous coding, all seven number items and eight other items were recoded, always by similarly combining adjacent codes, leaving only one number item – verbal counting – and three other items with values other than 0 or 1 (and a weight of 2).

We eliminated 37 items due to poor item fit and poor correlation with the scale, often supplemented by qualitative data. Fit indices ranging from .7 to 1.3 for the mean square (MNSQ) were conservatively preferred, and standardised MNSQ (ZSTD) of 0 ±2 were considered acceptable (Bond & Fox, 2001). Infit gives more weight to respondents close to an item's difficulty, whereas outfit is not weighted, giving relatively more influence to respondents not close to the item's difficulty level; therefore, infit values outside of these ranges were considered problematic. Item discrimination based on item–total score correlation was also examined, to ensure that all items contributed positively to the measurement of the latent trait.

Because no single fit statistic is sufficient to indicate item misfit, multiple fit statistics were examined, along with video analysis of student test behaviours, to determine whether items should be eliminated. As an example of an eliminated item, geometry item G02C ascertained

(a)



(b)

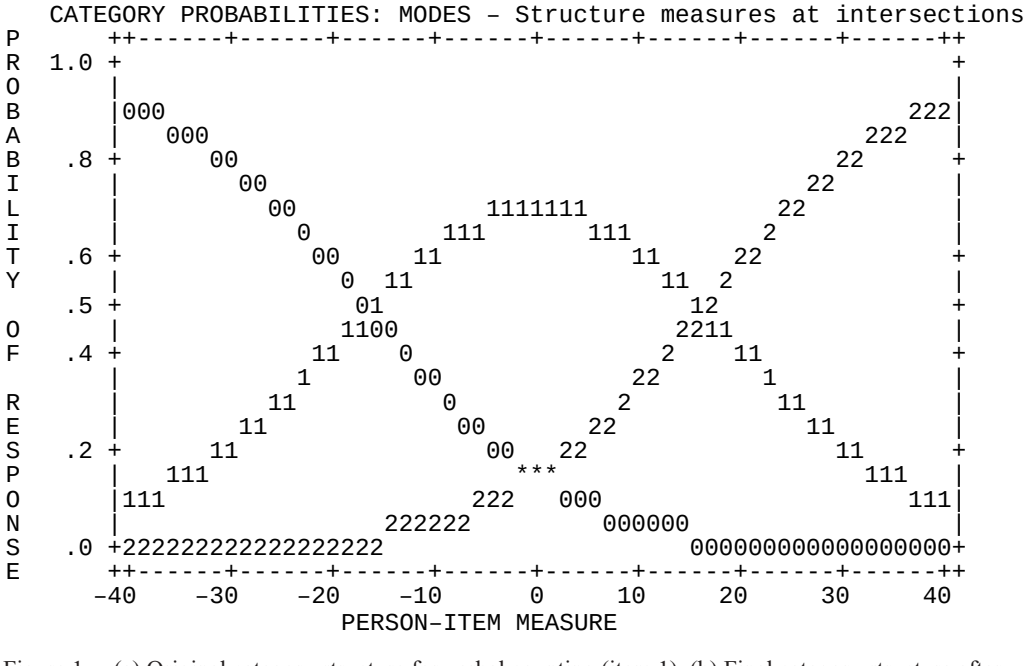


Figure 1. (a) Original category structure for verbal counting (item 1). (b) Final category structure after recoding.

whether children incorrectly matched (as congruent) two different size equilateral triangles. The correlation between this item and the scale score was negative. Further, this item's fit statistics were as follows: difficulty −1.66; infit MNSQ 1.45, ZSTD 9.9; and outfit MNSQ 1.72, ZSTD 9.9. Finally, qualitative observations of the assessment videos revealed that most children could not determine whether the two were different in size using visual or measurement strategies (e.g., after holding them up and inspecting them, some children declared they were the same size, even though they were not), supporting the item's elimination.

Other items similarly eliminated included two additional tasks about which the qualitative evidence indicated misinterpretations by children. For example, one item was a parallelogram in which not all sides were of equal length; however, the lengths were close, and several children perceived it to be a rhombus. Given that most of these children explicitly said that the sides were all the same, and that the item was designed to assess knowledge of the properties of geometric figures, not fine visual discrimination, the item was eliminated.

Several items were eliminated because they were redundant – that is, they measured the same level of a topic. For example, two items in number comparison, N34 ('Now I'm going to ask which is *smallest.* Which is *smallest*: 8 or 6 or 10?') and N52D ('What number is between 5 and 7?'), had almost the same difficulty (1.95 and 1.94). Although both were satisfactory items, only item N34, with slightly better fit statistics, was retained to reduce the length of the instrument's administration.

Retained items and item statistics

Of the 236 original items, 199 were retained. Table 1 provides the mathematics topics and examples of these retained items.

Table 2 reports the item difficulties (*t* scores), standard errors, infit and outfit statistics, partial correlations with the total measure, and score weights for each retained item. All had acceptable statistics in most categories; no item had all its fit statistics beyond acceptable ranges.

Some of the most difficult items had less satisfactory fit statistics, but so few children answered them that we retained them for future empirical testing. Similarly, a few items' fit statistics were slightly outside the ideal range, but qualitative observations generally revealed that children understood the items, and they were necessary for coverage of the levels of the particular developmental progression being assessed. In several cases, the less satisfactory fit statistics may have resulted from cultural influences. For example, neither fit nor correlations were satisfactory for items that involved selecting squares as rectangles, reflecting a common bias in the USA to reject such hierarchical classifications (Clements & Battista, 1992; Clements & Sarama, 2007b); however, they were retained for mathematical completeness and to detect potential positive effects of curricula on such misconceptions. (Recall that palpable distractors for the geometry items were weighted at only .05 of a point in the Rasch analysis. They were the least difficult items and are not included in the tables and figures for brevity's sake.)

A dimensionality analysis – that is, a factor analysis of standardised residuals – was computed to test the claim of unidimensionality. Loadings between −.3 and .3 indicate insignificant correlation between the items and any potential additional construct (Tabachnick & Fidell, 1983), supporting the assumption that the instrument measures only one construct in the Rasch model (Linacre, 2005). All items had loadings between −.2 and .2, with one exception – the group of palpable distractors for the shape identification items. Thus it is reasonable to assume that the items measure only one construct: early mathematics achievement.

Figure 2 presents the construct, or item and persons, map. The vertical axis is expressed in *t* scores. The person distribution is on the left, with the # symbol representing four children and the dot representing one to three children. Thus, children with the greatest ability in mathematics are

Table 1. Mathematics topic, number of items, and examples of retained item.

Topic	No. of items	Examples	
Comparing and ordering	21	N01 Ordinal	The bear, cat, and dog are waiting to go to the store to buy food. Which one is first in line?
		N17A Compare	Shown two cards, with 4 dots and 3 dots, child is asked, 'Which one has more?'
		N53 Compare	Which is closer to 590, 870 or 240?
Verbal counting	1	N11	How high can you count? Start at 1 and tell me.
Counting and counting strategies	25	N04	The child is shown 5 objects in a line and asked, 'bought these cans of food. Count these cans to tell me how many there are.'
		N27	Let's practice more counting. Please count to 10, starting at 4. [If child starts at 1, interrupt by saying] Please start at 4 and count to 10.
		N42	Here are 6 pennies. Three more are hidden under the cloth. How many are there in all?
		N57	Secretly hide 7 pennies under a cloth and put four on the table in a line. I have 11 pennies. Four are here, the rest are hidden under this cloth. How many are hidden?
		N61	There are 7 pennies under this cloth. There are 6 more under this cloth. How many are there in all [gesturing across both cloths]?
Arithmetic	10	N31	Provided manipulatives, the child is asked, 'Pretend I give you 3 candies and then I give you 2 more. How many will you have altogether?'
		N51	How much is 2 + 7? (no objects)
		N60	What is 47 + 25?
Recognition of number and subitizing	5	N03	The child is shown 2 small food cans on a picture of a shopping cart and given similar materials, then asked, 'Make yours look just like mine.'
		N40	Shown a card with 10 grapes for 2 seconds exactly, which is then hidden, the child is asked: 'How many?'
Composing number	5	N35	Look. I will put 4 boxes in this shopping cart. Count with me. 1, 2, 3, 4. 4! [Lay them in a straight line as you count.] Now, I'm going to hide some. [Cover the boxes with cloth, then secretly remove and hide 2, then remove the cover to show the remaining 2.] How many am I hiding?
		N59	I can make 6 with 3 and 3…. Show me a different way to make 6. [If the child is successful, say] Can you show me another way?
Geometry	6		
Comparing shape		G02C	Given 8 basic shapes, the child is asked to match shapes that are exactly the same shape and same size, while avoiding incorrectly matching palpably different shapes.
		G02B	Same task as G02C, but matching congruent triangles
		G16	Child must indicate that two 'L' shapes, one of which is slightly longer, will not exactly fit on each other.
Identifying shape	104	G03C	Asked to select all squares, child selects a large, prototypical square.

Table 1. *(Continued).*

Topic	No. of items	Examples	
		G05T	Asked to select all triangles, child correctly selects a non-prototypical triangle.
		G06C	Asked to select all rectangles, child correctly selects a square.
Turns	1	G13	Child completes an analogy involving rotated shapes.
Representing shape	2	G11	Child must make an accurate representation of a triangle using sticks.
Composing shape	8	G17B	Copies a design with a square inside a circle inside a square.
		G24	Fills 6 outlines of regular hexagons with pattern blocks, using different compositions for each.
Measuring	7	G01	Identifies which of 2 pictured pencils, not aligned, is the longer.
		G26	Measures a line segment with a single 1-inch strip of paper.
Patterning	6	G04	Identifies the missing element in the pattern ABA_AB.
		G18	Duplicates a single copy of the core unit of a pattern.

at the top of the scale. The distribution of children approximates a normal curve with an arithmetic mean of 50, indicating that the scale was appropriate for the participants. The items are on the right, with the easiest items at the bottom. The arithmetic mean of the items was a bit higher than that of the persons, which is appropriate because this assessment is intended to measure mathematical development of these children across several years. Figure 2 shows that the students' ability range is well covered by the item difficulty range; there is no noticeable gap in item difficulties among the items.

Finally, the item separation index, calculated as the proportion of variance accounted for by the Rasch model vs. the error variance, indicates how well the scale can differentiate items on the variable measured. The separation index of 6.66 indicates that this scale satisfactorily distinguished the items. Item reliability was .98 (KR-20 computed on the raw scores was a comparably high .94).

Descriptive statistics

Table 3 provides descriptive statistics for children in the autumn of their preschool year and toward the end of their spring semester. The children were in programmes that included mathematics curricula. The scores indicate, on average, growth in mathematics knowledge between autumn and spring of more than a standard deviation. In addition, consistent with previous research (Van de Rijt et al., 1999), there was a wide range of scores (about 48 for both age groups, or nearly five standard deviations on the *t* score).

Evaluation of developmental progressions

Figure 2 shows that the assumptions about developmental progressions were generally supported. That is, items measuring a given level in the developmental progression of each topic (defined *a priori*) were more difficult than items measuring lower levels and were less difficult than items measuring higher levels.

Table 2. Item difficulty, standard error, fit (infit and outfit, mean square residual and standardized), partial correlation, score weighting, and brief description for each retained item.

			In		Out				
Item	Difficulty	*SE*	MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Weight	Description*
N63	95.97	9.14	1.13	0.4	2.07	1.1	–0.13	1.00	How much smaller than 32 is 27?
G26	95.80	3.37	1.03	0.2	1.97	1.3	0.05	1.00	Measure segment*
N62	92.88	6.51	0.88	0	0.68	0.2	0.26	1.00	What is 4 more than 7?
N61	90.73	4.10	0.94	0	0.78	–0.2	0.25	1.00	Hide 7, show 6, how many?*
G25	89.04	2.42	1.01	0.1	2.79	2.8	0.06	1.00	Inverse size of unit vs. # of units
G24	88.21	2.26	1.01	0.1	1.89	1.7	0.10	1.00	Puzzle composition*
G23	87.54	6.75	1.21	0.5	1.74	0.9	0.10	1.00	Add 7 inch and 4 inch strips
N60	87.23	5.55	1.19	0.5	1.28	0.6	0.07	1.00	47 + 25*
N59	84.31	1.88	0.93	–0.3	0.46	–1.8	0.29	1.00	All partitions/compositions of 6*
G22	84.28	1.96	1.01	0.1	1.63	1.6	0.14	1.00	Inverse size of unit vs. # of units
N57	83.73	2.62	1.13	0.6	1.51	1.6	–0.02	1.00	11 in all, show 4, how many hidden?*
N58	82.85	11.07	0.49	–0.8	0.37	–0.6	0.87	1.00	69 + 3
G10C	81.58	1.64	1.21	1.4	5.78	7.9	–0.17	1	Select square as rhombus
N56	81.07	3.26	0.98	0	0.80	–0.2	0.32	1.00	How much is 6 + 4?
G10G	80.69	1.58	1.23	1.6	5.62	8.2	–0.19	0.25	Select square as rhombus
N55	80.60	2.09	1.15	0.9	1.59	1.6	0.08	1.00	6 – 4, comparison
G21	80.07	3.53	1.20	0.8	2.12	1.5	0.19	1.00	Measure with broken ruler
G18	80.03	1.63	1.09	0.7	1.38	1.3	0.15	1.00	Duplicate core unit of AB pattern*
N54	79.81	1.52	0.93	–0.5	1.00	0.1	0.28	1.00	5 + 3 with objects
N53	78.46	2.89	0.95	–0.2	0.90	–0.2	0.32	1.00	Which is closer to 590, 870 or 240?*
G06K	78.38	1.43	1.26	1.9	3.37	6	–0.10	0.25	Select square as rectangle
N49	78.06	1.69	1.01	0.1	1.07	0.4	0.22	1.00	What is 2 numbers after 7?*
G06C	77.92	1.40	1.28	2.1	5.00	8.8	–0.17	0.25	Select square as rectangle*
N51	77.47	3.22	0.78	–0.9	0.54	–0.9	0.50	1.00	2 + 7*
G06G	77.06	1.35	1.30	2.4	5.14	9.4	–0.20	0.25	Select square as rectangle
N52	77.03	1.40	0.83	–1.5	0.51	–2.8	0.44	1.00	Insert cube tower

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
N50	76.94	1.35	1.21	1.7	2.59	4.9	0.01	1.00	4 + _ = 9
G19	76.42	1.32	0.97	–0.2	1.51	2.1	0.25	1.00	Find strip measured by 4 inch strips
G10K	75.68	1.28	1.19	1.7	2.84	5.9	0.03	0.25	Select square as rhombus
N48	75.37	2.01	0.91	–0.6	1.05	0.3	0.37	1.00	4 + 3
N46	74.65	2.30	1.10	0.7	1.41	1.2	0.22	1.00	4 + _ = 6 (composition)
N45	74.21	1.25	0.85	–1.5	0.73	–1.6	0.42	1.00	Order cube towers
N44	74.20	5.43	0.65	–1.6	0.57	–1.4	0.75	1.00	How many in 10 by 4 array?
N47	73.70	6.60	1.05	0.3	2.23	1.1	0.34	1.00	6 + _ = 10 (composition)
N43	73.44	8.88	0.48	–1.3	0.45	–1.3	0.91	1.00	58 + 1
N42	73.22	1.53	0.91	–0.9	0.89	–0.8	0.39	1.00	6 shown, 3 hidden, how many?*
N41	72.75	2.13	0.98	–0.1	0.92	–0.4	0.33	1.00	Which is smaller, 32 or 27?
N39	72.20	1.44	1.05	0.6	1.16	1.2	0.19	1.00	9 – 4 with objects
N40	71.13	2.27	0.98	–0.1	0.99	0.1	0.35	1.00	Subitize 10*
G17C	70.72	1.07	0.93	–0.9	0.88	–0.8	0.38	1.00	Compose transparent shapes
N38	70.50	1.10	0.82	–2.4	0.59	–3.3	0.50	1.00	Order sets of 1–6
N37	68.99	1.10	0.99	–0.1	0.87	–1.2	0.36	1.00	Count 30 objects
N35	67.94	1.13	1.08	1	1.37	2.2	0.22	1.00	Show 4, hide 2, how many hiding?*
G20	67.75	2.39	1.38	2.9	1.66	2.7	–0.06	1.00	2 + _ = 4 (composition)
N34	67.58	1.51	0.97	–0.4	1.06	0.6	0.35	1.00	Duplicate core unit of ABC pattern
N36	67.09	0.96	1.03	0.5	1.10	0.8	0.34	1.00	2 + _ = 5 (composition)
N32	65.77	1.19	0.92	–1.5	0.88	–1.7	0.42	1.00	Number before 8
G17A	65.52	0.92	0.96	–0.6	1.04	0.4	0.40	1.00	Compose transparent shapes
N33	65.27	0.91	0.85	–2.7	0.64	–3.7	0.52	1.00	Order numerals 1–5
N30	64.94	1.42	0.92	–1.5	0.93	–0.7	0.43	1.00	Biggest, 7, 9, 5
N31	64.74	0.90	1.02	0.3	1.14	1.3	0.36	1.00	3 + 2*
G15	63.61	0.88	1.05	1	1.21	2	0.35	1.00	Cardinality, counting 8 objects
N29	62.89	1.58	1.15	2.7	1.32	3.6	0.13	1.00	Which is closer to 6, 9 or 4?

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
N28	62.77	1.09	0.91	–2	0.87	–2.2	0.45	1.00	Count backwards from 10
G14	61.64	0.85	0.81	–4.4	0.65	–4.5	0.58	1.00	Make ABB pattern
N26	60.42	0.92	0.95	–1.3	0.94	–1.1	0.43	1.00	Count 15 pennies
G16	60.33	0.82	1.37	7.5	1.64	6.7	0.12	1.00	Congruence – rectangles
G13	60.29	0.82	1.16	3.4	1.32	3.7	0.29	1.00	Turn analogy*
N25	59.85	1.10	1.10	2.1	1.19	2.5	0.30	1.00	Which is more 9 large or 11 small objects
N27	59.63	0.83	0.89	–2.7	0.90	–1.5	0.51	1.00	Verbally count from 4*
G12	59.26	0.59	0.96	–0.7	0.93	–1.4	0.58	2.00	Build a rectangle with sticks
N23	58.88	0.89	0.91	–2.4	0.89	–1.9	0.48	1.00	Duplicate 6 objects
G05Y	58.64	0.80	1.27	6	1.53	6.4	0.21	0.50	Identify non-prototypical triangle
N24	58.08	1.36	0.86	–2.8	0.84	–2.2	0.52	1.00	Biggest, 5, 6, 4
G10I	57.64	0.79	1.18	4.4	1.33	4.5	0.29	0.50	Select rhombus
G11	56.84	0.56	1.05	1	1.06	1	0.55	2.00	Build triangle with sticks*
G10Z	56.31	0.78	1.26	6.3	1.34	4.9	0.25	0.50	Select prototypical rhombus
N06DD	56.06	1.44	1.28	3.8	1.40	3.4	0.24	1.00	Subitize 6
N22	56.04	0.86	1.05	1.3	1.02	0.5	0.37	1.00	Identify counting mistake – cardinality
G07C	55.57	0.77	0.93	–1.8	0.85	–2.6	0.52	1.00	Puzzle composition
G10A	55.36	0.77	1.24	5.9	1.33	5.2	0.27	0.50	Select prototypical rhombus
G07D	55.23	0.77	0.91	–2.4	0.84	–2.9	0.53	1.00	Puzzle composition
G02B	54.98	0.77	1.10	2.6	1.13	2.1	0.38	1.00	Match triangles*
G09	54.97	0.77	0.81	–5.4	0.76	–4.6	0.60	1.00	Puzzle composition
G05T	54.89	0.77	1.35	8.5	1.51	7.7	0.18	0.50	Identify non-prototypical triangle*
G05U	53.96	0.76	1.23	5.9	1.32	5.3	0.28	0.50	Identify non-prototypical triangle
G7B	53.89	0.76	0.89	–3	0.82	–3.4	0.55	1.00	Puzzle composition
N19	53.77	1.26	1.08	1.5	1.10	1.3	0.21	1.00	Produce set of 10
N17D	53.23	0.76	0.68	–9.9	0.59	–9	0.71	1.00	Match numeral to set – 4
N20	53.17	0.77	0.80	–6.2	0.74	–5.8	0.62	1.00	Cardinality – 8

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
N17E	52.83	0.76	0.67	–9.9	0.59	–9.3	0.72	1.00	Match numeral to set – 4
G05J	52.39	0.75	1.43	9.9	1.62	9.9	0.12	0.50	Identify non-prototypical triangle
G06R	51.55	0.75	1.00	0.1	0.98	–0.4	0.47	0.50	Identify prototypical rectangle
G08	51.14	0.76	0.81	–5.7	0.75	–5.5	0.61	1.00	Make AB pattern
N17C	51.01	0.75	0.65	–9.9	0.58	–9.8	0.73	1.00	Match numeral to set – 3
G07A	50.63	0.65	0.83	–3.4	0.83	–3.5	0.63	2.00	Puzzle composition
G05N	50.26	0.75	1.31	8.1	1.44	7.7	0.22	0.25	Identify prototypical triangle
N18	49.95	1.06	0.86	–3	0.80	–2.8	0.52	1.00	Produce set of 7
G06X	48.91	0.75	1.03	1	1.03	0.6	0.44	0.50	Identify rectangle
N17B	48.88	0.75	0.64	–9.9	0.57	–9.9	0.74	1.00	Match numeral to set – 2
N14	48.30	0.75	0.95	–1.4	0.93	–1.4	0.50	1.00	Duplicate 4 objects
N16	48.03	0.86	1.02	0.6	1.08	1.3	0.38	1.00	Match numeral to set – 1
N17A	47.97	0.76	0.63	–9.9	0.56	–9.9	0.75	1.00	Which is more, 4 or 3?*
G05D	47.95	0.75	1.16	4.4	1.23	4.2	0.33	0.25	Identify prototypical triangle
N15	47.85	0.78	0.76	–7.2	0.71	–6.1	0.64	1.00	Produce a set of 5
N013	47.71	0.76	0.94	–1.6	0.93	–1.2	0.50	1.00	Cardinality – 5
G03G	47.63	0.75	1.14	3.7	1.25	4.4	0.34	0.25	Select prototypical square
G03K	45.73	0.76	1.36	8.9	1.62	9.2	0.15	0.50	Select square
G04	45.64	0.84	0.99	–0.3	0.92	–1.3	0.47	1.00	Find missing pattern element*
G17B	45.45	0.88	1.01	0.3	0.99	–0.2	0.43	1.00	Copy design*
G05H	45.07	0.76	1.77	9.9	2.35	9.9	–0.19	0.25	Avoid non-triangle
N11	44.90	0.66	0.79	–4.4	0.78	–4.6	0.66	2.00	Verbal counting from 1*
G05L	44.60	0.77	1.63	9.9	2.05	9.9	–0.08	0.50	Avoid non-triangle
N12	44.54	0.77	0.95	–1.4	1.01	0.2	0.48	1.00	Count 8 objects
G02A	43.93	0.78	1.04	1	1.03	0.5	0.42	1.00	Match squares
N10	43.82	0.78	1.12	3.1	1.06	0.9	0.37	1.00	Compare 3 objects to 3 objects
G03C	42.48	0.78	0.97	–0.8	1.05	0.8	0.45	0.25	Select prototypical square

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
N09	42.38	1.08	1.10	1.5	1.18	1.7	0.26	1.00	Count 8 objects in a line
G02D	41.49	0.80	1.29	6.6	1.42	5.1	0.20	1.00	Avoid mistakes in matching shapes
N05	41.07	0.92	1.05	1	1.30	3	0.31	1.00	Count 4 objects
G03Q	40.47	0.81	1.27	6	1.73	7.7	0.17	0.25	Avoid non-closed ‘square’
G01	39.80	0.82	1.23	4.9	1.35	3.8	0.23	1.00	Which is longer?*
G02C	39.46	0.82	1.21	4.5	1.50	5.2	0.22	1.00	Match shapes*
N07	38.42	0.84	0.91	−1.9	0.83	−1.9	0.48	1.00	Compare 3 objects to 4 objects
G06V	38.08	0.84	1.56	9.9	2.64	9.9	−0.13	0.25	Avoid non-rectangular parallelogram
N03	37.30	0.86	0.97	−0.6	0.89	−1.2	0.43	1.00	Duplicate 2 objects
N05	34.58	0.92	0.89	−2	0.80	−1.7	0.46	1.00	Count 4
N04	33.10	0.95	0.93	−1.2	1.03	0.3	0.40	1.00	Count 5 objects*
G10F	33.06	0.95	1.29	4.4	1.93	5.8	0.08	0.05	Avoid hexagon as rhombus
N03	31.94	0.98	0.94	−1	0.80	−1.5	0.41	1.00	Duplicate 2 objects*
G10W	31.90	0.98	1.27	3.8	2.02	5.8	0.07	0.05	Avoid pentagon as rhombus
G10S	31.68	0.98	1.29	4.1	2.10	6.1	0.05	0.05	Avoid octagon as rhombus
G03O	31.57	0.99	0.96	−0.5	0.84	−1.1	0.38	0.05	Avoid circle as square
G03N	31.12	1.00	1.05	0.7	1.16	1.1	0.29	0.05	Avoid triangle as square
N01	30.93	1.14	0.87	−1.5	0.63	−2.5	0.43	1.00	Who is first?*
N02	30.50	1.35	0.97	−0.2	0.93	−0.3	0.30	1.00	Which is more, 3 or 4 objects?
G03J	30.42	1.02	1.04	0.6	1.05	0.4	0.29	0.05	Avoid palpable distractor selecting shapes
G03P	30.06	1.03	1.02	0.3	0.93	−0.4	0.32	0.05	″
G03Z	29.31	1.06	1.18	2.2	1.69	3.6	0.13	0.25	″
G03L	29.18	1.06	1.06	0.7	1.09	0.6	0.27	0.05	″
G06F	29.18	1.06	1.12	1.5	1.42	2.3	0.20	0.05	″
G06M	29.18	1.06	1.24	2.9	2.01	4.8	0.07	0.05	″
G03M	28.79	1.08	1.02	0.3	1.09	0.6	0.29	0.05	″
G03I	28.65	1.08	1.14	1.7	1.68	3.4	0.17	0.25	″

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
G06W	28.52	1.09	1.12	1.5	1.32	1.8	0.19	0.05	″
G10O	28.52	1.09	1.07	0.9	1.12	0.7	0.25	0.05	″
G06S	27.97	1.11	1.15	1.7	1.40	2.1	0.16	0.05	″
G06E	27.39	1.13	1.19	2.1	1.71	3.3	0.10	0.05	″
G06Z	27.39	1.13	1.13	1.5	1.42	2.1	0.16	0.05	″
G10V	27.39	1.13	1.22	2.4	2.52	6	0.03	0.05	″
G03H	26.79	1.16	1.07	0.8	1.34	1.7	0.22	0.05	″
G06Y	26.48	1.17	1.12	1.3	1.55	2.5	0.15	0.05	″
G10Y	26.17	1.19	1.14	1.5	1.69	3	0.13	0.05	″
G10P	25.84	1.20	1.09	1	1.72	3	0.16	0.05	″
G03F	24.99	1.24	1.03	0.3	0.93	–0.3	0.26	0.05	″
G03X	24.81	1.25	1.09	0.9	1.65	2.6	0.14	0.05	″
G10E	24.81	1.25	1.15	1.4	1.55	2.3	0.11	0.05	″
G06A	24.45	1.27	1.14	1.3	1.62	2.5	0.11	0.05	″
G06T	24.45	1.27	1.10	0.9	1.49	2	0.15	0.05	″
G06U	24.27	1.28	1.09	0.8	1.30	1.3	0.17	0.05	″
G05G	24.08	1.29	1.06	0.5	1.26	1.2	0.20	0.05	″
G10Q	24.08	1.29	1.06	0.6	1.29	1.3	0.19	0.05	″
G03Y	23.89	1.30	1.13	1.2	1.79	2.9	0.10	0.05	″
G05P	23.89	1.30	1.11	1	2.04	3.6	0.11	0.05	″
G06O	23.89	1.30	1.04	0.4	1.08	0.4	0.22	0.05	″
G06Q	23.89	1.30	1.07	0.7	1.53	2.1	0.16	0.05	″
G03D	23.50	1.32	1.10	0.9	1.49	1.9	0.15	0.05	″
G06H	23.30	1.33	1.06	0.5	1.16	0.7	0.20	0.05	″
G06I	23.10	1.34	1.14	1.2	1.88	3.1	0.08	0.05	″
G03U	22.89	1.35	1.11	1	1.68	2.5	0.12	0.05	″
G05K	22.89	1.35	1.06	0.5	1.25	1.1	0.18	0.05	″

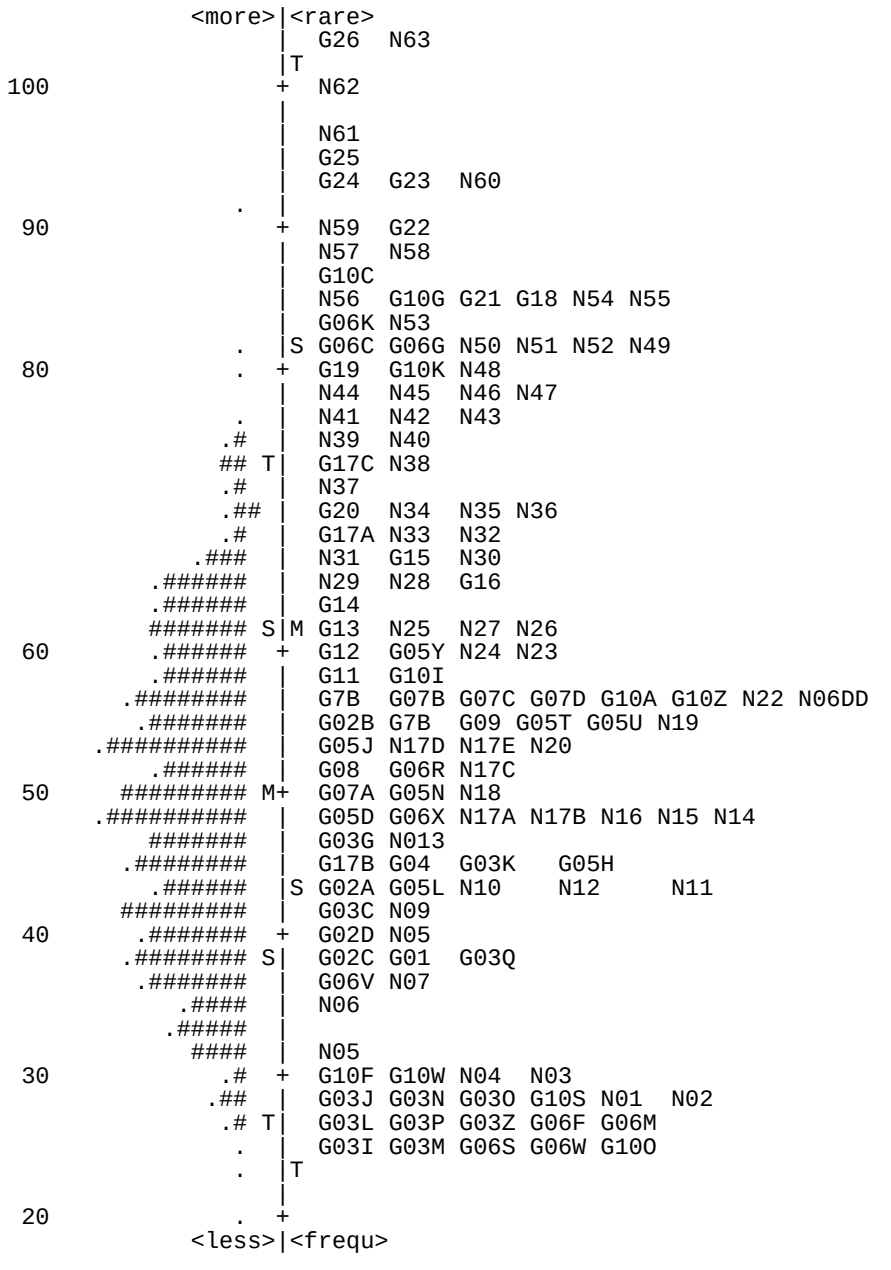
Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
G05M	22.89	1.35	1.09	0.8	1.69	2.5	0.14	0.05	″
G06N	22.89	1.35	1.06	0.6	1.12	0.5	0.19	0.05	″
G10H	22.89	1.35	1.06	0.5	1.11	0.5	0.19	0.05	″
G10X	22.89	1.35	1.04	0.4	1.00	0.1	0.22	0.05	″
G03V	22.68	1.36	1.09	0.8	1.57	2.1	0.12	0.05	″
G06J	22.68	1.36	1.08	0.7	1.57	2.1	0.13	0.05	″
G03T	22.46	1.38	1.09	0.8	1.62	2.2	0.13	0.05	″
G05Q	22.46	1.38	1.00	0	1.21	0.9	0.23	0.05	″
G05C	22.25	1.39	1.06	0.6	1.22	0.9	0.17	0.05	″
G05V	22.25	1.39	1.10	0.8	3.14	5.7	0.07	0.05	″
G06L	22.25	1.39	1.09	0.8	1.54	2	0.13	0.05	″
G03E	22.02	1.40	1.05	0.5	1.35	1.3	0.18	0.05	″
G03S	22.02	1.40	1.04	0.4	1.08	0.4	0.20	0.05	″
G05X	22.02	1.40	1.06	0.5	1.96	3.1	0.15	0.05	″
G06D	22.02	1.40	1.05	0.4	1.05	0.3	0.20	0.05	″
G03W	21.80	1.42	1.07	0.6	1.30	1.2	0.16	0.05	″
G10L	21.56	1.43	1.10	0.8	1.73	2.4	0.11	0.05	″
G06B	21.33	1.45	1.07	0.6	1.29	1.1	0.15	0.05	″
G10U	21.33	1.45	1.13	1	1.81	2.6	0.07	0.05	″
G03B	21.09	1.46	1.06	0.5	1.62	2.1	0.15	0.05	″
G06P	21.09	1.46	1.07	0.5	1.66	2.2	0.13	0.05	″
G03A	20.84	1.48	1.09	0.7	1.39	1.4	0.13	0.05	″
G05F	20.84	1.48	1.01	0.1	0.91	–0.3	0.23	0.05	″
G05Z	20.84	1.48	1.06	0.5	1.71	2.3	0.14	0.05	″
G05E	20.07	1.53	1.02	0.2	1.07	0.4	0.19	0.05	″
G05I	19.80	1.55	1.04	0.3	1.45	1.5	0.16	0.05	″
G10T	19.80	1.55	1.06	0.5	1.17	0.7	0.15	0.05	″

Table 2. *(Continued)*.

Item	Difficulty	*SE*	In		Out		Corr.	Weight	Description*
			MNSQ	ZSTD	MNSQ	ZSTD			
G03R	19.52	1.57	1.10	0.7	2.18	3.1	0.06	0.05	″
G05O	19.52	1.57	0.98	–0.1	1.13	0.5	0.22	0.05	″
G10B	19.52	1.57	1.08	0.6	1.62	1.9	0.11	0.05	″
G10J	19.52	1.57	1.02	0.2	1.18	0.7	0.18	0.05	″
G10N	19.24	1.59	1.03	0.3	1.15	0.6	0.18	0.05	″
G05A	18.65	1.63	1.05	0.4	1.31	1	0.14	0.05	″
G10R	18.65	1.63	1.03	0.2	1.07	0.3	0.18	0.05	″
G05R	17.69	1.71	1.05	0.4	2.49	3.4	0.10	0.05	″
G05W	17.35	1.73	1.03	0.3	1.44	1.3	0.14	0.05	″
G10D	17.35	1.73	1.03	0.3	1.24	0.8	0.15	0.05	″
G05S	16.64	1.79	1.00	0	0.82	–0.4	0.20	0.05	″
Mean	60.13	1.68	0.99	–0.3	1.15	0			
SD	17.09	1.8	0.18	3.5	0.62	3.6	0	0	0

Note: * indicates a more detailed description is available in Table 1.



Note: # represents four children. Most of the palpable distractors from the shape identification items (GSHP…) have been removed for brevity's sake.

Figure 2. Person–item map.

Table 3. Descriptive statistics of *t*-scores by age.

Age	Minimum	Maximum	Mean	*SD*
4.25	20.25	69.08	44.42	7.85
5.00	34.79	83.51	56.15	8.49

For example, the easiest items in the number comparison (NCRN) progression were the simplest ordinal task (naming the 'first' object, N01) and simplest number comparison (comparing small sets of items non-verbally, N02). Somewhat more difficult were two other number comparison tasks, involving visual comparison of slightly larger sets. The next five items involved verbally-based comparison of small numbers. Counting-based comparison of larger sets (e.g., N25) was more difficult. The remainder of the items assess the development of extensive mental number lines and ordering sets and measures. The hypothesised progression was supported, with one significant departure. We hypothesised that children's ability to order would develop after the development of the mental number line up to 10. However, the results indicated that, rather than a clearcut progression from mental number lines to ordering, different ordering tasks were interwoven with the mental number line tasks in the following sequence: ordering sets of up to five dots in canonical arrangements (N33), ordering up to six pictures in random arrangements (N38), and ordering lengths of connected cubes (6, 7, 8, 9, 11 12; N45) or inserting a missing length (10) into that series (N52). The developmental progression theory and the ordering of items on the instrument were altered to fit these data.

Other progressions were similarly supported. The object counting and strategies progression began with the ability to maintain correspondence when counting, with small linear sets (N04) first, then with increasing larger linear sets (N05, N05, N09), and then with sets not in a line (N12). Next is the ability to state that the last counting number represents the numerosity of the set (the cardinality principle). Subsequent developments include the ability to produce a small set (of five), to recognise obvious errors such as skipping an object (N16), to apply the cardinality principle to sets of 10 and more (N20), and to produce sets of more than five (N18 and N19). Children can then recognise more complex counting errors such as misapplying the cardinal principle (N22), and can start counting at numbers other than one (N27), count unorganised arrangements of 10 and more (N26), and count backward (N28). Later they can name numbers immediately before or after other numbers (N32), count objects above 30 (N37), and use increasingly sophisticated counting strategies to solve arithmetic tasks (N39 'How much is 9 take away 4?'; N63 'How much smaller is 27 than 32?'). The arithmetic progression is, of course, closely related to the counting strategies. The tasks were discriminated by their verbal reference to real-world situations modeled by arithmetic: for example, the union of two disjoint sets of concrete objects. Most items fit the progression, with the exception of the first multiplication items, which were easier than predicted (however, note the caveat that few children responded to these and other difficult items).

The remainder of the number progressions included fewer items, and were simpler in their construction, usually proceeding from smaller to larger numbers. These were supported without exception, including the recognition of number and subitising and the composition of number progressions.

Other domains show similar results. The geometry items dealing with shape recognition, although simple tasks, tap into developmental progressions that are relatively difficult to trace. All the difficult items were at the level at which children had to identify properties of classes of geometric shapes; for example, recognising a square as a (special type of) rectangle. The next most difficult items were on levels at which children had to select atypical examples of familiar

shapes (e.g., an obtuse, 'skinny' triangle rotated from the horizontal; G05Y, G10I) or exemplars of less familiar shapes (e.g., a rhombus with a vertical longer diagonal – the prototypical 'diamond'; G10A). Easier again were the recognition of examples of familiar shapes that are slight variants (e.g., an otherwise prototypical triangle without a horizontal base; G05J) or exemplars (e.g., a prototypical rectangle with a horizontal base; G06R, G05N, G05D), or, similarly, avoiding the incorrect selection of distractors (e.g., chevrons are triangular shapes with curved sides; G05H). Finally, the easiest items were the many palpable distractors (e.g., a circle as a distractor for a square or rhombus).

The results from the shape composition items mirrored research confirming developmental progression (Clements, Wilson et al., 2004). Items measuring simple matching to outlines (G17B) were easier than items in which children could use side length and other clues in trial-and-error strategies (G07A), which were easier than intentionally composing shapes to make other shapes (G07B, G07D, and G17C are designed to measure both of these levels; G17A and G17C measure just the latter), which was easier than conscious substitution of shapes within compositions (G24).

Items assessing shape matching and comparison corresponded exactly to the proposed developmental progression, with the three easiest items (G02C, G02D, G02A) reflecting matching of identical shapes, the next easiest (G02B) assessing ability to match congruent shapes in different orientations, and next were two others that did not allow manipulation of the shapes, at the levels of comparing on the basis of gestalt similarity (G15) and on the basis of explicit analysis of shapes' attributes (G16). The remaining geometry developmental progressions had too few items to justify such analyses.

Items assessing length measurement were mostly difficult, except for one at the lowest level, in which children compared lengths using a third length (G01). As hypothesised, the items on the higher levels were considerably more difficult. Results on an item requiring respondents to measure by placing multiple copies of a unit end-to-end (G19), on two requiring measurement with a ruler (G21, G22), and on the single 'conceptual ruler' (Steffe, 1991) item (e.g., ability to analyse length relationships arithmetically; G23) were in the hypothesised order. In contrast, two items expected to precede length measurement, involving understanding the inverse relationship between unit size and number of units (G25), and measuring by iterating a single unit (G26), were the most difficult measurement items. This may be due to the low number of respondents, inadequacies of the tasks, or inaccuracies as regards the higher levels of the developmental progression. A larger number of respondents, especially older children, must be assessed before there is sufficient data on which to base a decision.

Our earlier research had indicated that, contrary to our initial hypothesis, identifying the missing element in a simple (ABAB) pattern was easier even than duplicating or extending patterns. This latest data collection bore that out, and supported the hypothesised developmental progression, with that task (G04) easier than duplicating AB patterns (G08), extending AB patterns (G09), extending more sophisticated patterns (G14), and recognising the core unit of a pattern (G20 and G18).

Discussion

The Research-Based Early Maths Assessment (REMA) uses an individual interview format, with explicit protocol, coding, and scoring procedures. It assesses children's thinking and learning along research-based developmental progressions for topics in mathematics considered significant for preschoolers, as determined by a consensus of participants in a national conference on early childhood mathematics standards (Clements et al., 2004).

After using the data to revise the coding of 15 items, all items had a satisfactory category structure. Thus, the codes for all items represented distinct categories and discriminated children's

responses well. Elimination of 37 items that either showed poor fit and poor correlation with the scale, or were redundant, resulted in the inclusion of 199 items that measured different levels within each developmental progression and that adequately fit the Rasch model (recall that four geometry tasks involved 104 items).

These data provided empirical support for the proposed developmental progression within most topics. The majority (seven) were supported with no modification; minor modifications made to two developmental progressions brought theory and data into alignment. Only one trajectory, measurement, needs additional revision and testing before its higher levels can be reliably described. (The remaining three topics – verbal counting, geometric transformations, and representing shapes – included only one or two items.)

The construct map (Figure 2) approximated a normal curve with an arithmetic mean of 50. Therefore, this assessment and scale was appropriate for the population of children for which it was designed. The arithmetic mean of the items was a bit higher than that of the persons. Considering that the instrument is intended to measure the knowledge of children up to the traditional second grade level, and thus had many items at the more difficult end of the scale, this is understandable.

In summary, the final instrument satisfactorily measured mathematics ability as a coherent, unidimensional latent trait and possessed the properties of conjoint measurement, producing interval measures that are linear and additive. Use of the Rasch model provides strong indications that the measured behaviours are expressions of the underlying construct of mathematics ability. Thus, there is support for the instrument's construct validity.

Conclusions and implications

The REMA, along with other similar efforts (Thomson et al., 2005), fills a gap in the measurement of mathematics achievement in the early years (NICHD Forum, Washington, DC, June 2002; CIRCL Forum, Temple University, January, 2003). First, the REMA provides a measure of the mathematical topics deemed important by mathematicians, mathematics and early childhood educators and educational researchers, and teachers, as well as those included on most state and national standards. Second, this comprehensive measure has been submitted to, and validated by, the Rasch model. Construct validity in the Rasch model is a comprehensive concept that includes content validity, face validity, and concurrent validity (Bond & Fox, 2001; Smith, 2004), and thus the work undertaken for this study complements the procedures employed in previous formative assessments to establish the content and concurrent validity of the instrument. Rasch modeling is a theory-based approach to developing measures through hypothesis testing (Andrich, 2004; Wilson, 2005); when data fit the Rasch model, this is evidence of the construct validity of the instrument (Smith, 2004). The validity of our original hypothesis of the unidimensional progression of early mathematical achievement is thereby supported.

Thus, the final version of the REMA has the essential properties of measurement, linearity, and additivity, as well as high reliability and validity as traditionally conceptualised. Equal interval scales are important for any evaluation or research study. For example, the common approach to measuring change uses gain scores; however, such gain scores are typically negatively correlated with initial level, especially when not using equal-interval scales, so that children with low pretest scores artificially show the largest gain. Further, such a scale is more robust in the treatment of omitted or guessed items – especially important in the assessment of young children, for whom behavioural and attention issues may strongly impact responses to any item or set of items. Omitted items in classical testing are treated as if the child had answered incorrectly. In Rasch assessment, such items do not negatively impact the determination of the child's score. Similarly, guessing has less impact too; the model does not give credit for correct answers to difficult items from low ability students.

Together, its psychometric properties imply that the REMA can be a useful measure of mathematics achievement among preschoolers. Given the increasing numbers of children attending early care and education programmes, increasing recognition of the importance of mathematics, and increasing interest in evaluating and studying pre-kindergarten programmes designed to facilitate academic achievement, instruments that assess children on a common ability scale with a consistent, justifiable metric will play a critical role, with benefits for clinical assessment, professional knowledge, pedagogy, and curricula. The descriptive and qualitative information that the test also provides, not the focus of the present analysis but documented in our previous studies (Clements & Sarama, 2004; Sarama & Clements, 2002), is another advantage of this approach to assessment.

In addition, this study provides data from a linear, equal interval scale on fine-grained theoretically- and empirically-based developmental progressions in major topics in early childhood mathematics. The results generally confirmed the validity of these developmental progressions, which confirms the usefulness of the instrument, and also provide critical information for researchers interested in each of the topics assessed, curriculum developers, and teacher educators (Fuson et al., 1997; Griffin, 2004; National Council of Teachers of Mathematics, 2006), as regards international research on developmental progressions in early mathematics (e.g., Canobi, Reeve, & Pattison, 2002; B. Clarke & Shinn, 2004; D. M. Clarke et al., 2002; Mulligan, Mitchelmore, & Prescott, 2005; Mulligan, Prescott, Papic, & Mitchelmore, 2006; Siegler & Booth, 2004; Van de Rijt & Van Luit, 1999; Yuzawa & Bart, 1999). The few cases in which the hypothesised developmental progression was altered or questioned constitute particularly important areas for future research. Such research might include additional items on other areas such as fractions and partitioning, the role of units and unitising across topics, and the role of pattern and structure across topics (Mulligan et al., 2005, 2006).

There are several caveats. The final instrument and the developmental progressions must be evaluated again, as adjustments were made to each after data collection. More importantly, we could not ascertain how children's prior experiences may have affected their responses to the items and thus their relative difficulties (Thomson et al., 2005). Results from research using multiple methodologies will have to be synthesised to address such complex issues. Also, this study is limited to three- to four-year-old children. We are presently collecting additional data to address these limitations, including extending participation to children up to eight years of age. In addition, certain topics that some programmes or curricula include, even for very young children, such as probability, fractions, time, and money, are not measured. We consider this a strength; the REMA measures multiple topics, but only those determined by research and a consensus of informed parties (Clements et al., 2004) to be appropriate and important for young children. Thus, we have avoided the unfortunate proliferation of objectives in too many standards and curricula, even as we developed a comprehensive measure of mathematics. Others, however, may believe that a useful topic has been neglected. Another limitation is the lack of information about the REMA's prediction of achievement longitudinally. We are collecting such data as well.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143–166). Maple Grove, MN: JAM Press.

Barnett, W.S., Hustedt, J.T., Hawkinson, L.E., & Robin, K.B. (2006). *The state of preschool 2006.* New Brunswick, NJ: National Institute for Early Education Research.

Baroody, A.J. (2004). The developmental bases for early childhood number and operations standards. In D.H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 173–219). Mahwah, NJ: Lawrence Erlbaum Associates.

Blevins-Knabe, B., & Musun-Miller, L. (1996). Number use at home by children and their parents and its relationship to early mathematical performance. *Early Development and Parenting, 5,* 35–45.

Bond, T.G., & Fox, C.M. (2001). *Apply the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bracken, B.A. (1998). *Bracken Basic Concept Scale–Revised.* San Antonio, TX: Psychological Corporation, Harcourt Brace and Company. (Original work published 1984)

Canobi, K.H., Reeve, R.A., & Pattison, P.E. (2002). Young children's understanding of addition concepts. *Educational Psychology, 22,* 513–532.

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.

Clarke, D.M., Cheeseman, J., Gervasoni, A., Gronn, D., Horne, M., McDonough, A., et al. (2002). *Early Numeracy Research Project final report.* Melbourne: Department of Education, Employment, and Training, the Catholic Education Office, and the Association of Independent Schools Victoria.

Clements, D.H., & Battista, M.T. (1992). Geometry and spatial reasoning. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 420–464). New York: Macmillan.

Clements, D.H., & Conference Working Group. (2004). Part one: Major themes and recommendations. In D. H. Clements, J. Sarama & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 1–72). Mahwah, NJ: Lawrence Erlbaum Associates.

Clements, D.H., & Sarama, J. (2004). Building Blocks for early childhood mathematics. *Early Childhood Research Quarterly, 19,* 181–189.

Clements, D.H., & Sarama, J. (2007a). Building Blocks—SRA Real Math Teacher's Edition, Grade PreK. Columbus, OH: SRA/McGraw-Hill.

Clements, D.H., & Sarama, J. (2007b). Early childhood mathematics learning. In F.K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 461–555). New York: Information Age Publishing.

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

Clements, D.H., Sarama, J., & DiBiase, A.-M. (2004). *Engaging young children in mathematics: Standards for early childhood mathematics education.* Mahwah, NJ: Lawrence Erlbaum Associates.

Clements, D.H., Sarama, J., & Gerber, S. (2007). Mathematics knowledge of low-income entering preschoolers. Manuscript submitted for publication.

Clements, D.H., Sarama, J., & Wilson, D.C. (2001). Composition of geometric figures. In M.v.d. Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 273–280). Utrecht, The Netherlands: Freudenthal Institute.

Clements, D.H., Wilson, D.C., & Sarama, J. (2004). Young children's composition of geometric figures: A learning trajectory. *Mathematical Thinking and Learning, 6,* 163–184.

de Lemos, M., & Doig, B. (1999). *Who am I? Developmental assessment manual.* Camberwell, Victoria, Australia: Australian Council for Educational Research.

Doig, B., & de Lemos, M. (2000). *I can do maths.* Camberwell, Victoria, Australia: Australian Council for Educational Research.

Doig, B., McCrae, B., & Rowe, K. (2003). *A good start to numeracy: Effective numeracy strategies from research and practice in early childhood.* Canberra: Australian Council for Educational Research.

Fuson, K.C. (1997). Research-based mathematics curricula: New educational goals require programs of four interacting levels of research. *Issues in Education, 3,* 67–79.

Fuson, K.C., Smith, S.T., & Lo Cicero, A. (1997). Supporting Latino first graders' ten-structured thinking in urban classrooms. *Journal for Research in Mathematics Education, 28,* 738–760.

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).

Griffin, S. (2004). Number Worlds: A research-based mathematics program for young children. In D.H. Clements, J. Sarama, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 325–342). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffin, S., Case, R., & Capodilupo, A. (1995). Teaching for understanding: The importance of the central conceptual structures in the elementary mathematics curriculum. In A. McKeough, J. Lupart, & A. Marini (Eds.), *Teaching for transfer: Fostering generalization in learning* (pp. 121–151). Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkle, D. (2000). *School involvement in early childhood.* Washington, DC: National Institute on Early Childhood Development and Education, US Department of Education Office of Educational Research and Improvement.

Holloway, S.D., Rambaud, M.F., Fuller, B., & Eggers-Pierola, C. (1995). What is "appropriate practice" at home and in child care? Low-income mothers' views on preparing their children for school. *Early Childhood Research Quarterly, 10,* 451–473.

Jordan, N.C., Huttenlocher, J., & Levine, S.C. (1992). Differential calculation abilities in young children from middle- and low-income families. *Developmental Psychology, 28,* 644–653.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Klein, A., Starkey, P., & Wakeley, A. (2000). *Child Math Assessment: Preschool battery (CMA).* Berkeley: University of California.

Linacre, J.M. (2005). *A user's guide to Winsteps/Ministep Rasch-model computer program.* Chicago: Winsteps.com.

Mulligan, J., Mitchelmore, M., & Prescott, A. (2005). Case studies of children's development of structure in early mathematics: A two-year longitudinal study. In H.L. Chick & J.L. Vincent (Eds.), *Proceedings of the 29th Conference of the International Group for Psychology in Mathematics Education* (Vol. 4*,* pp. 1–8). Melbourne, Australia: PME.

Mulligan, J., Prescott, A., Papic, M., & Mitchelmore, M. (2006). Improving early numeracy through a Pattern and Structure Mathematics Awareness Program (PASMAP). In P. Clarkson, A. Downton, D. Gronn, A. McDonough, R. Pierce, & A. E. Roche (Eds.), *Building connections: Theory, research and practice (Proceedings of the 28th annual conference of the Mathematics Education Research Group of Australia)* (pp. 376–383). Sydney, Australia: Mathematics Education Research Group of Australia.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

National Center for Education Statistics. (2000). *America's kindergartners (NCES 2000070).* Washington, DC: U.S. Government Printing Office.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

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

Sarama, J., & Clements, D.H. (2002). Building Blocks for young children's mathematical development. *Journal of Educational Computing Research, 27*(1&2), 93–110.

Sarama, J., Clements, D.H., & Vukelic, E.B. (1996). The role of a computer manipulative in fostering specific psychological/mathematical processes. In E. Jakubowski, D. Watkins & H. Biske (Eds.), *Proceedings of the 18th annual meeting of the North America Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 567–572). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Saxe, G.B., Guberman, S.R., & Gearhart, M. (1987). Social processes in early number development. *Monographs of the Society for Research in Child Development, 52*(2, serial no. 216).

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

Smith, R.M. (2004). Fit analysis in latent trait measurement models. In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Starkey, P., Klein, A., Chang, I., Qi, D., Lijuan, P., & Yang, Z. (1999). *Environmental supports for young children's mathematical development in China and the United States.* Albuquerque, NM: Society for Research in Child Development.

Steffe, L.P. (1991). Operations that generate quantity. *Learning and Individual Differences, 3,* 61–82.

Tabachnick, B.G., & Fidell, L.S. (1983). *Using multivariate statistics.* New York: Harper & Row.

Thomson, S., Rowe, K., Underwood, C., & Peck, R. (2005). *Numeracy in the early years: Project Good Start.* Camberwell, Victoria, Australia: Australian Council for Educational Research.

Van de Rijt, B.A.M., & Van Luit, J.E.H. (1999). Milestones in the development of infant numeracy. *Scandinavian Journal of Psychology, 40,* 65–71.

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

Watson, J.M., Callingham, R.A., & Kelly, B.A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9,* 83–130.
Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock–Johnson III Tests of Achievement* (3rd ed.). Itasca, IL: Riverside Publishing Company.
Wright, B.D., & Stone, M.H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.
Wright, R.J. (2003). A mathematics recovery: Program of intervention in early number learning. *Australian Journal of Learning Disabilities, 8*(4), 6–11.
Yuzawa, M., & Bart, W.M. (1999). *Development of size comparison strategies in young children.* Albuquerque, NM: Society for Research in Child Development.