Article

Journal of Psychoeducational Assessment
2021, Vol. 39(8) 983–998

DOI: 10.1177/07342829211037195
journals.sagepub.com/home/jpa

Measuring Early Childhood Mathematical Cognition: Validating and Equating Two Forms of the Research-Based Early Mathematics Assessment

Yixiao Dong[1], Douglas H. Clements[1], Crystal A. Day-Hess[1], Julie Sarama[1], and Denis Dumas[1]

Abstract

Psychometric work with young children faces the particular challenge that children's attention spans are relatively short, and therefore, shorter assessments are required while retaining comprehensive coverage. This article reports on three empirical studies that encompass the development and validation of the research-based early mathematics assessment-short form (REMA-SF), an instrument that measures the early mathematical competency of children from 3 to 8 years of age. The developed measure captures both children's mathematical performance and the strategies children use to solve math problems. Results indicated that the REMA-SF can produce valid scores for measuring children's math skills in early childhood, and the validity of the measure can be well-generalized to an external (or independent) sample. Additionally, we also equated the REMA scores between the long and short forms of the assessment: anchor items common across the forms were selected and refined in the equating process.

Keywords

early mathematics, construct validity, cross-observer validity, test equating

Introduction

A burgeoning interest in early mathematics has accentuated the paucity of assessment instruments for young children, especially those that provide a comprehensive and consistent assessment from preschool through the primary grades. However, conducting psychometric work with young children faces the particular challenge that children's attention spans are relatively short, and

[1]University of Denver, Denver, Colorado, USA

Corresponding Author:
Douglas H. Clements, University of Denver, 1999 E. Evans Ave, Denver, Colorado, USA.
Email: Douglas.Clements@du.edu

therefore, shorter assessments are required while retaining comprehensive coverage. We report three empirical studies that encompass the development and validation of a short form of an early math assessment.

Need for Early Math Assessments

There are at least three reasons that the need for high-quality, effective math assessments is critical. First, there is increased recognition of the importance of mathematics and the varying performance of students in different countries in mathematics, beginning in the preschool years (e.g., Clements & Sarama, 2021; OECD, 2014). Second, these early years are surprisingly important for development through life. What math children know when they enter kindergarten predicts their later math achievement, reading achievement, and even high school graduation and socioeconomic status at age 42 (Duncan & Magnuson, 2011; Watts et al., 2014), suggesting the foundational role of mathematical thinking in cognitive development (Clements & Sarama, 2011). Third, early interventions in mathematics can prevent later learning difficulties in school for all children (e.g., Clements & Sarama, 2021; Fuson et al., 1997). However, advancements in both research and practice have been restrained by the lack of high-quality assessments of early math cognition. That is, extending our knowledge of young children's mathematical development and evaluating the effectiveness of programs designed for them requires accurate measures of their mathematical knowledge and skill.

Available instruments are useful for specific purposes but are limited in several ways. For example, a common instrument is the Woodcock–Johnson III (Woodcock et al., 2001), particularly the Applied Problems section. This subtest has several strengths, including assessing a wide range of abilities and ages, reliabilities above .80, and large normative data samples. However, two national panels on preschool assessment (NICHD Forum, Washington, DC, June 2002; CIRCL Forum, Temple University, January 30–31, 2003) cautioned sole use of the Woodcock–Johnson for assessment of mathematical skills in preschoolers. Specifically, the panel stated that it has not been validated for children in the youngest age ranges; covers a narrow range of problems; jumps quickly to advanced, formal knowledge; and is not based on current research on the development of mathematical thinking. As another example, the Bracken Basic Concept Scale (Bracken, 1984–1998) includes several mathematical concept categories; however, the national panels cautioned that content validity was low, and it is difficult to administer or interpret results for mathematical topics.

Challenges and Opportunities in Early Childhood Measurement

Admittedly, some limitations in existing instruments result from the challenges of assessing young children. A primary constraint is young children's short attention spans and their teachers' shared desire for them to be out of the classroom for testing for minimal periods (Clements et al., 2021; Shepard, 1994). Another challenge of assessing young children is that interaction with an assessor is necessary for a developmentally appropriate and valid measure of correctness and strategy. However, the intensity of such interactions and young children's distractibility can lead to the incorrect administration of items and is often costly.

Exacerbating these challenges is the increasing recognition of young children's competence in wide-ranging mathematical topics (at least 20 topics, Clements & Sarama, 2021), creating pressure for longer measures. In addition, researchers have documented various interesting strategic processes young children can use to solve math problems and have provided evidence that the development of such practices is as consequential as content knowledge (Clements et al., 2020). Children's solution strategies in mathematics are a particularly critical component of their

learning in that domain (e.g., Biddlecomb & Carr, 2011; Clements & Sarama, 2021; Siegler, 1993). Children's invented strategies may contribute to accuracy, problem-solving ability, base-10 number concepts, and flexibility in transferring knowledge in arithmetic (Carpenter et al., 1998). A final challenge is measuring mathematical competencies along the increasingly well-researched developmental progressions within each of these many topics longitudinally along a continuum of ages with an equal-interval scale. A measure that is capable of creating an equal-interval scale across ages would be important for longitudinal comparisons. Equal intervals along the continuum are superior for research in that gains scores, most especially when not using equal interval skills, are typically correlated with the initial level of performance, with the result that children within low pre-test scores artificially show the greatest gains.

Fortunately, researchers from various fields have conducted extensive research on children's mathematical thinking, including, recently, an extended number of topics (see reviews in Clements & Sarama, 2021; Sarama & Clements, 2009). Moreover, some assessments have been developed that meet at least some of the following criteria that were determined to be critical (Clements et al., 2021): (1) developmentally appropriate administration, including assessment time being as short as possible while meeting all other criteria; (2) robust even with some misadministered or skipped items; (3) coverage of a range of critical mathematical topics; (4) assessment, scoring, and coding of both correctness and strategy; (5) measure mathematical competencies along with the increasingly well-researched developmental progressions within each of these many topics and to do so along a continuum of ages with an equal-interval scale; and (6) psychometrically sound reliability and validity of scores.

REMA: A Validated Early Math Measure

The research-based early mathematics assessment (REMA, Clements et al., 2008) is an individual interview assessment based on scripted protocols of nearly 200 items designed to measure children's mathematical cognition from three to 8 years of age. The REMA uses a flipbook of pictures and manipulatives, which meets criterion 1 of developmentally appropriate administration and setting (albeit not the time aspect). In addition, REMA scores have been validated using a Rasch modeling methodology (Clements et al., 2008), and a modest number of missing items, due to misadministration or skipping, have been shown not to impact the determination of the child's score negatively. Similarly, there is less impact for guessing, thus addressing criterion 2.

Addressing criterion 3, the REMA provides a measure of the multiple mathematical topics deemed important by mathematicians, researchers, teachers, and state and national standards. Topics in number include verbal counting, object counting, subitizing, number comparison, number sequencing, connection of numerals to quantities, number composition and decomposition, adding and subtracting, multiplying and dividing, and place value. Geometry topics include shape recognition, shape composition and decomposition, congruence, construction of shapes, spatial imagery, as well as geometric measurement, patterning, and reasoning.

Beyond correctness, the REMA also collects, codes, and scores children's strategies when those are observable and relevant (e.g., processes of perceptual subitizing, or quick recognition of the number of objects in a set, would be neither). These strategies are recoded into four levels of sophistication for modeling, thus meeting criterion 4. Finally, addressing criterion 5, the REMA resulted from years of research and development of trajectories of development of mathematical cognition and the topic-specific developmental progressions (Clements et al., 2011, 2013; Sarama et al., 2012).

Finally, addressing the psychometrics of criterion 6, construct validity in the Rasch model is a comprehensive concept that includes content validity, face validity, and concurrent validity (Bond & Fox, 2015; Smith, 2004), complementing the procedures employed in the previous formative

assessments to establish the content and concurrent validity of the instrument (Clements et al., 2008). In addition, because Rasch modeling is a theory-based approach to developing measures through hypothesis testing (Andrich, 2004; Wilson, 2005), when data fit the Rasch model, there can be evidence of the construct validity of the instrument (Smith, 2004), supporting the unidimensional progression of early mathematical achievement.

As noted, the REMA (hereafter the REMA-Full) failed to address the second part of criterion 1—short assessment time. Not only are the youngest children's attention spans limited, but the main request from teachers and administrators is that children be out of the classroom for testing only for minimal periods. Even more pressure to decrease time of administration has come from research colleagues, who face both these pressures from school personnel and funding constraints: individual interviews are time-intensive, as is training assessors on hundreds of items, and thus expensive. As such, decreasing training and administration time was the primary motivation to develop the REMA-short form (REMA-SF; Clements et al., 2017).

Developmental Process and Content of the REMA-SF

As with the REMA-Full, the REMA-SF is intended to measure the mathematical competency of children from 3 to 8 years as a latent trait via item response theory (IRT). Items of the REMA-SF were chosen from the REMA-Full item pool (197 items). The item selection for the REMA-SF followed two criteria. First, we systematically retained content coverage as the full version but with fewer items within each topic, and items spread out along the difficulty levels as evenly as possible. Second, we reviewed and evaluated the item function based on the data of REMA-Full (e.g., TRIAD data, Clements et al., 2011) and selected those with better psychometric properties. The initial REMA-SF consisted of 80 items ordered by the Rasch difficulty parameters from previous studies (e.g., Clements et al., 2008), 50 in number and arithmetic, and 30 from other topics (i.e., patterning and algebraic thinking, shapes, shape composition, and measurement). In addition, we made revisions from multiple aspects based on small pilot samples, including item order, coding schemes for correctness and strategy, item phrasing, and assessment materials.

The REMA-SF was administrated via individual interviews using a standardized protocol. All assessors received a series of training sessions (i.e., orientation, demonstration, and practice) and a certification session to ensure standardized delivery. Although the REMA-SF has much fewer items than its full version, it still can be stressful and time consuming for young children to answer all 80 items. Therefore, administration of this test includes a start and stop rule. Children begin the assessment at the designated start points for each grade level. A floor or basal level must be established in which children get at least three consecutive items correct. All children, regardless of their grade level, stop the assessment after getting three contiguous questions incorrect.

The rating scale of REMA item correctness varies across items. Specifically, most items use dichotomous coding (0 = incorrect and 1 = correct), but some items also come with a partially correct code (0 = incorrect, 1 = partially correct, and 2 = correct). In addition to these correctness codes, many items also collect information on children's observable strategies for solving the math problems (e.g., a child counted on from six or four to find out "how much is 6 + 4"). These strategy codes are later recoded into different levels of sophistication of thinking. Both correctness and strategic sophistication indicators are incorporated into the Rasch scoring, resulting in 133 scoring variables.

Current Research

The present research aims to investigate the psychometric properties of the REMA-SF and equate scores from the two forms of the REMA. These general goals can be expressed in terms of three specific empirical research questions:

1. Do psychometric statistics support the validity of the shortened form?
2. Does the REMA-SF have cross-observer validity in an independent sample?
3. How can the scores from the REMA-SF be equated with REMA-Full scores?

The work is divided into three studies to answer these questions. The first study focused on the validity evidence (e.g., content and construct validity) of the REMA-SF. In the second study, the psychometric properties of the short form were evaluated using an independent sample. In the last study, the REMA scores from the long (full) and short forms of the REMA were equated via selected anchor items. Moreover, Smith et al. (2000) identified nine specific methodological "sins" in the literature on developing short-form tests, and these sins have been widely used as a framework to evaluate the quality of developed short measures (e.g., Dong et al., 2020; McCrae & Costa, 2007). The current study also employs their framework to check whether those common pitfalls were avoided in developing the REMA-SF.

Study One: Validating the REMA-SF

Methodology

The main sample was derived from a DREME Network longitudinal research study of the coherence of early childhood mathematics, which was supported by the Heising-Simons Foundation. Participants were children from PreK to grade 2 in two large, public school districts in the Western United States. The main sample consisted of longitudinal early mathematical data collected at six different time points: Spring and Fall 2017, 2018, and 2019. A total of 164 PreK children participated in the pre-test (i.e., Spring 2017). Additional children were recruited in the later waves of data collection, which resulted in 1912 observations (54% female). In addition, a cross-sectional sample with 218 children was collected in Spring 2016 for the initial pilot stage.

Following the methodology for developing the REMA-Full version (Clements et al., 2008), this study also applied the Rasch techniques. The construct validity of the REMA-SF was examined using the main sample. It should be noticed that the construct validity in Rasch is a comprehensive concept that involves content validity, face validity, and concurrent validity (Bond & Fox, 2015). Because the assessment contains both dichotomous and polytomous items, partial-credit Rasch models were conducted to generate the main psychometric statistics and evidence, including dimensionality, fit, construct coverage, and reliability and separation. All Rasch analyses were performed in *Winsteps 4.6* (Linacre, 2021).

Results

Dimensionality. Principal components analysis of residuals (PCAR) was tested to evaluate the dimensionality of the REMA-SF. The measure explained 42.9% of the variance with the first contrast eigenvalue of 4.1 with 1.8% unexplained variance, which means the first dimension explained a large portion of the total raw variance (over 40%, Linacre, 2016). In contrast, an extra dimension can only explain a negligible portion of the variance. This finding supports the unidimensionality of the REMA-SF.

Reliability and Separation. We checked both person reliability and separation index to examine the reproducibility of relative measure location from the REMA-SF. The Rasch person reliability was equivalent to the classical "test" reliability, while a separation value was the number of statistically different performance strata that the test can identify in the sample (Linacre, 2016). The REMA-SF showed person reliability of .93 and separation of 3.58 (larger than 2.0 is often recommended, Wright & Masters, 1982). Thus, both statistics indicated that REMA-SF produced highly reliable scores.

Model-Data Fit. Mean square (MNSQ) fit statistics were used to evaluate the quality of the REMA-SF measure and items. In general, mean squares near 1.0 indicate little distortion of the measurement system (Linacre, 2002). The overall *MNSQ* of the test was 1.02 (*SD* = .38). At the item level, according to the recommended item fit range (.6 to 1.4) by Wright and Linacre (1994), four of 133 (3%) scoring indicators were identified as misfitting items. However, these items are still retained in the current REMA-SF for some theoretical reasons, principally that those items reflected important early mathematical content.

Furthermore, because of the start and stop rules of the REMA, a few very difficult items did not receive enough responses to generate robust estimates. For example, the last item of the test (strategies to complete the number sentence "75+47+25 = _") showed the largest misfitting value (*Infit MNSQ* = 2.01) as well as the largest standard error (SE = 1.32). Such items were included in the current REMA-SF for assessing high achievers that may appear in the future although they temporarily demonstrated misfitting because of the limited responses.

Construct Coverage. The item–person (Wright) map in Figure 1 informed us of the targeting and construct coverage of the REMA-SF. The person ability (left) and item difficulty (right) estimates were distributed along the ability/difficulty continuum. As can be seen, items were well-targeted to the examinees, and the item difficulties almost fully covered the range of examinees' abilities, which indicates REMA-SF was able to measure children with various ability levels in this sample. Some REMA items located at the same logit positions (i.e., same difficulty levels) could be used for further item deletions, but they were all retained here to support content coverage and content validity.

In conclusion, the content coverage of the original REMA-Full was systematically retained in its short form, and the psychometric statistics supported the validity of the shortened REMA.

Study Two: Cross-Observer Validity of the REMA-SF

A major concern of any newly developed test is whether it can show validity in an independent sample other than the one with which it was initially developed and validated (Smith et al., 2000). Thus, study two was conducted to examine such cross-observer validity of the established REMA-SF.

Methodology

This external sample was derived from a site in the eastern United States that partnered with us to train their staff to administer the REMA. The data consisted of 101 kindergarteners (51% female) and 80 first graders (54% female). All children in the external sample took the current version of the REMA-SF at both pre- and post-tests in the fall and spring, respectively. We investigated the cross-observer validity of the REMA-SF, that is, whether the validity of the measure can be generalized to examinees from other groups or situations. The Rasch analyses conducted in study

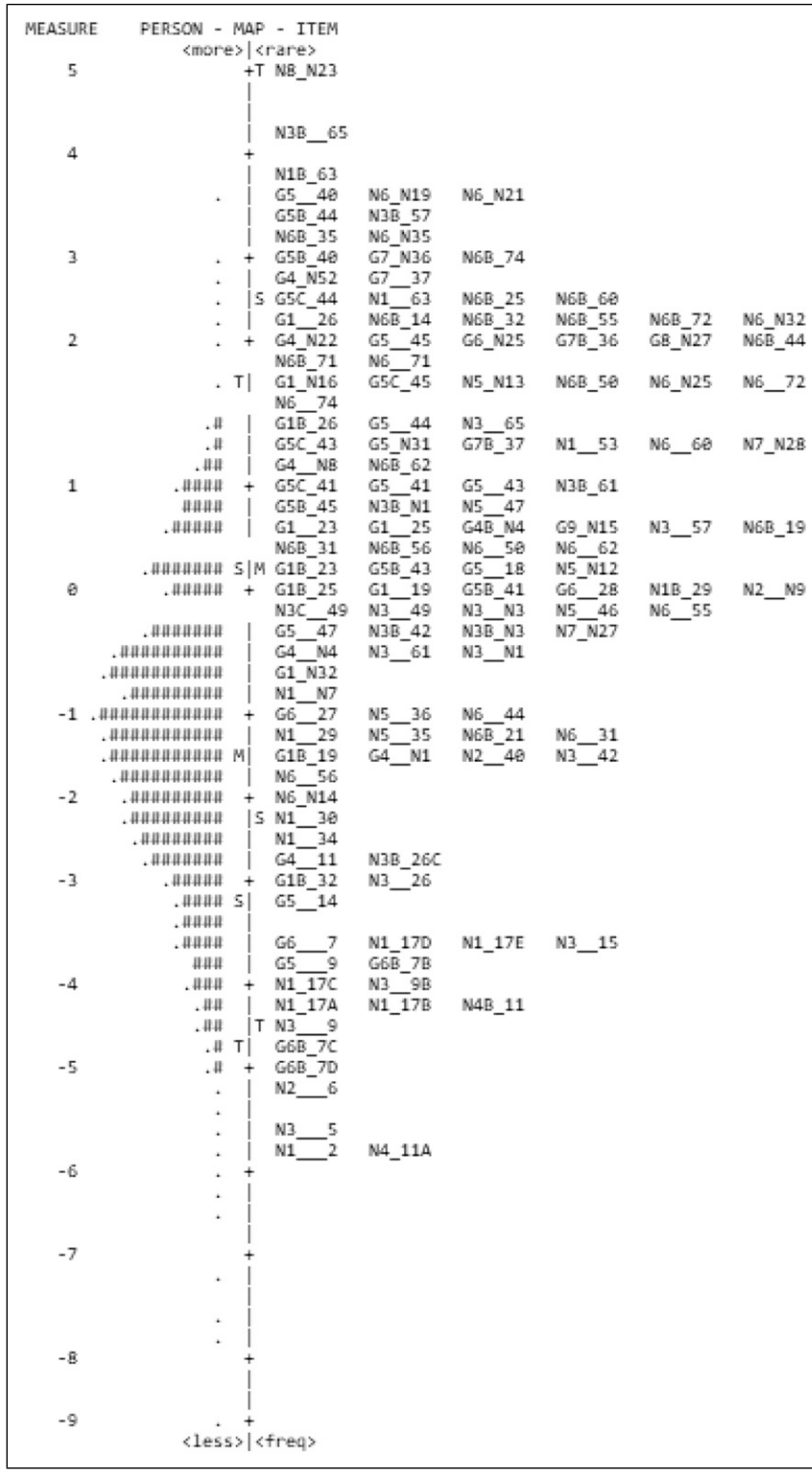

Figure 1. Item–person (Wright) map of the REMA-SF with the main sample. Notes. Each "#" is 10, and each "." is 1 to 9.

one, including dimensionality, fit, construct coverage, and reliability and separation, were repeated with the data from study two.

Results

The findings in study two are reported here in a summary manner to reduce redundant interpretations. From the PCAR results, the measure explained 40% of the variance with the first contrast eigenvalue of 4.1, and an extra dimension can only explain an additional 3% variance, so the REMA-SF was still able to demonstrate a unidimensional structure within the independent dataset used here. The REMA-SF also produced highly reliable scores (*person reliability* = .90 and *separation* = 3.06). Moreover, the overall mean square of the test was still near 1.0 (*MNSQ* = .99, *SD* = .35), and there was only one item showing substantial misfitting to the model. From Figure 2, the REMA-SF displayed good construct coverage and targeted the external sample participants.

In conclusion, the cross-observer validity of the REMA-SF was supported because the validity evidence found in study one was replicated in the independently collected external sample.

Study Three: Score Equating between Two Forms of the REMA

Methodology

Study three used an additional sample in which all participants took the full version of the REMA. The REMA-Full data were from a large-scale randomized control trial study (Clements et al., 2011), which included 1305 children (51% female; age ranging from 44 to 64 months in the PreK year, M = 52.06, SD = 4.09) from two large urban school districts in the Northeastern United States.

Using this REMA-Full sample and the REMA-SF sample in study one**,** score equating analyses were conducted between the two forms of the REMA. Specifically, the Common-Item Equating (CIE) approach, the most highly recommended extant Rasch equating method (Linacre, 2021), was applied in the current research. We chose correctness indicators (as opposed to strategic processing indicators) as anchor items in the analysis for several reasons. Since the REMA-Full was first developed (Clements et al., 2008), there were fewer modifications in the correctness codes than the strategy codes, so a substantial number of correctness items have the same content in the two REMA forms. Additionally, those items spread throughout the difficulty continuum, an ideal characteristic for anchors (Dorans et al., 2010).

Results

The prerequisite of equating two tests is that both have been validated in separate analyses (Dorans et al., 2010; Linacre, 2021). Studies one and two have presented considerable validity evidence of the REMA-SF, and the REMA-Full was validated in previous work (Clements et al., 2008). Therefore, we proceeded to equate the scores between the two versions of the REMA.

Initial CIE. In the initial analysis, all correctness indicators were used as the anchors. Table S-1 in the online supplemental materials presents the item difficulties of all common items in two forms of the REMA and statistics for testing the differences among them. Using the 95% critical values (i.e., $|t| \leq 1.96$), a total of 29 items demonstrated non-significant differences in item difficulties between two forms of the REMA. Notably, the difficulty of these items spread across the entire ability continuum, which indicates they can collectively be ideal anchors for the equating purpose.

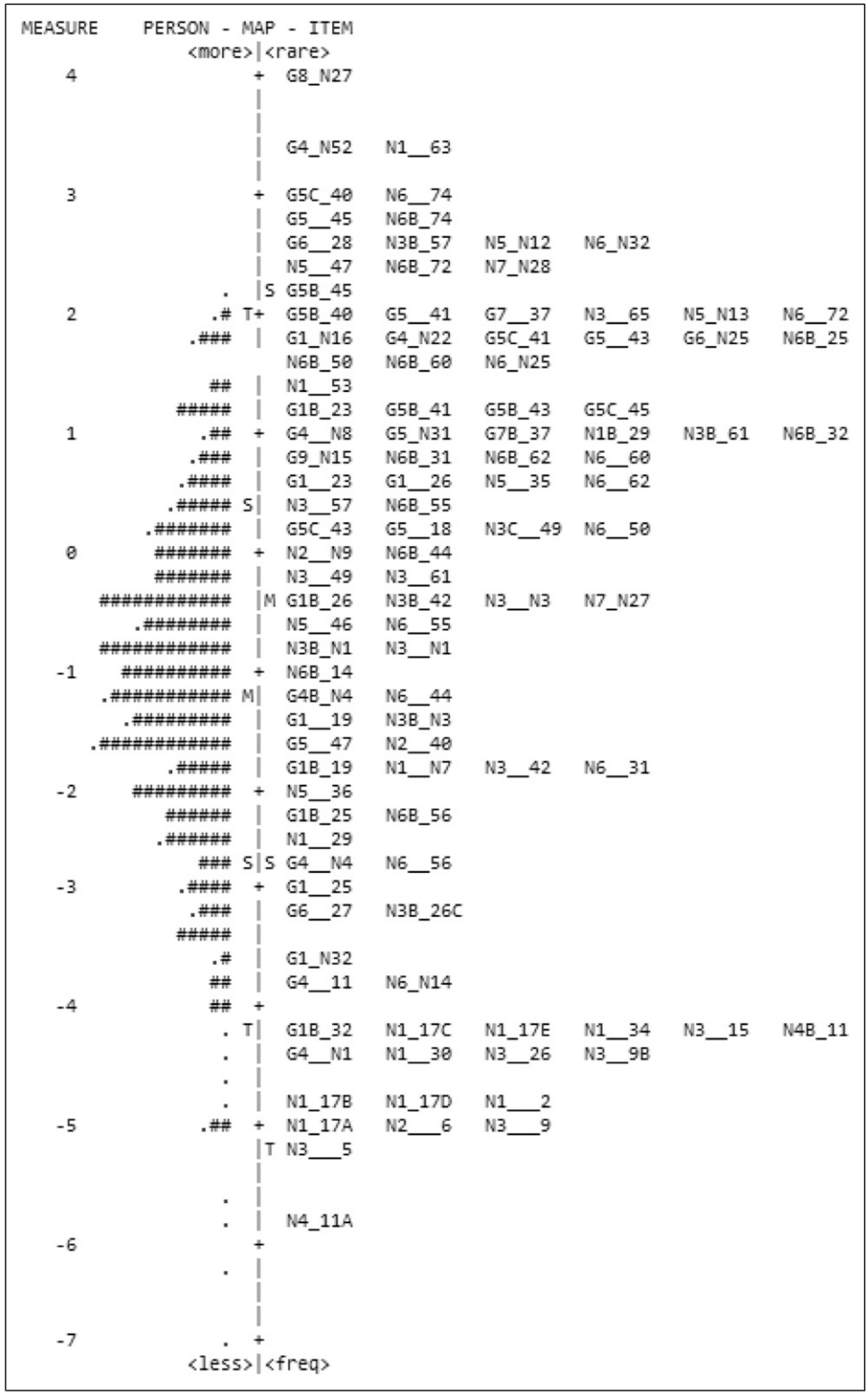

Figure 2. Item–person (Wright) map of the REMA-SF with the long branch sample. Notes. Each "#" is 2, and each "." is 1.

The difficulties of the common items from the two REMA forms are also visually depicted in Figure 3(a). Appropriate anchors are expected to be located near the empirical line (i.e., the dotted line) and within confidence bands (i.e., two solid curves). However, some items (e.g., 1, 54, and 83) substantially deviated from the zone. The slope of this line was .95, and Linacre (2021) suggested a slope value near 1.0 to perform a feasible score approximation between the tests via the following function

$$\theta' = \theta - \overline{\theta}_C + \overline{\theta'}_C \quad (1)$$

In equation (1), θ represents the measures from the REMA-SF, and θ' is the corresponding measures in the REMA-Full frame of reference; $\overline{\theta}_C$ is the mean of common items from the REMA-SF; and $\overline{\theta'}_C$ is the mean of common items from the REMA-Full. Although the current slope estimate was not far from the best-fit slope (1.0), it was still necessary to conduct the equating by the most qualified anchor items. Therefore, we conducted a second CIE using the 29 refined common items with similar difficulties across two forms of the REMA.

CIE with the Refined Common Items. Table 1 displays the results of the second CIE analysis. Only one of the selected common items showed significant differences in difficulty between the REMA-SF ($\theta = -2.66$) and REMA-Full ($\theta' = -2.14$), $t = -2.57$, $p < .05$. From Figure 3(b), almost all anchor items were very close to the empirical line with a slope of .98. The findings indicated that these items behaved in the same way across two forms of the REMA and were appropriate for the use of equating REMA scores.

In conclusion, REMA scores from the short form can be reasonably equated with the scores from the full version via the 29 selected anchor items. Given that the averaged θ of the REMA-SF common items was .26 ($SD = 2.17$) and the averaged θ' of the REMA-Full common items was .58 ($SD = 2.12$), the approximate REMA-SF measures (θ) of the REMA-SF sample in the REMA-Full frame of reference (θ') can be calculated via the function (1): $\theta' = \theta - (.26) + (.58) = \theta + .32$.

General Discussion

Assessing young children's mathematical cognition has been a challenging task. One major constraint is young children's short attention spans and their teachers' shared desire for short assessment time. Despite analyses showing that the original REMA-Full is a well-validated instrument (Clements et al., 2008), the goal of short and timely assessment time was not met with the original measure. We were therefore motivated to develop and validate the REMA-Short Form in the present study.

Is the REMA-SF a "High Quality Short Measure"?

To better understand the quality of the developed REMA-SF, we applied Smith et al.'s (2000) framework for evaluating short instruments to examine the present work. We found that the methodological procedures involved in developing the REMA-SF have met the suggestions of creating a high-quality short assessment.

Specifically, the REMA-SF has preserved the content coverage of the original REMA. In the development of the short form, we pursued the balance between solid psychometric properties and sufficient content coverage (e.g., a few items were retained though they showed misfitting). Meanwhile, the unidimensional structure of the REMA-Full was reproduced in its short form. According to the sufficient reliability and separation values

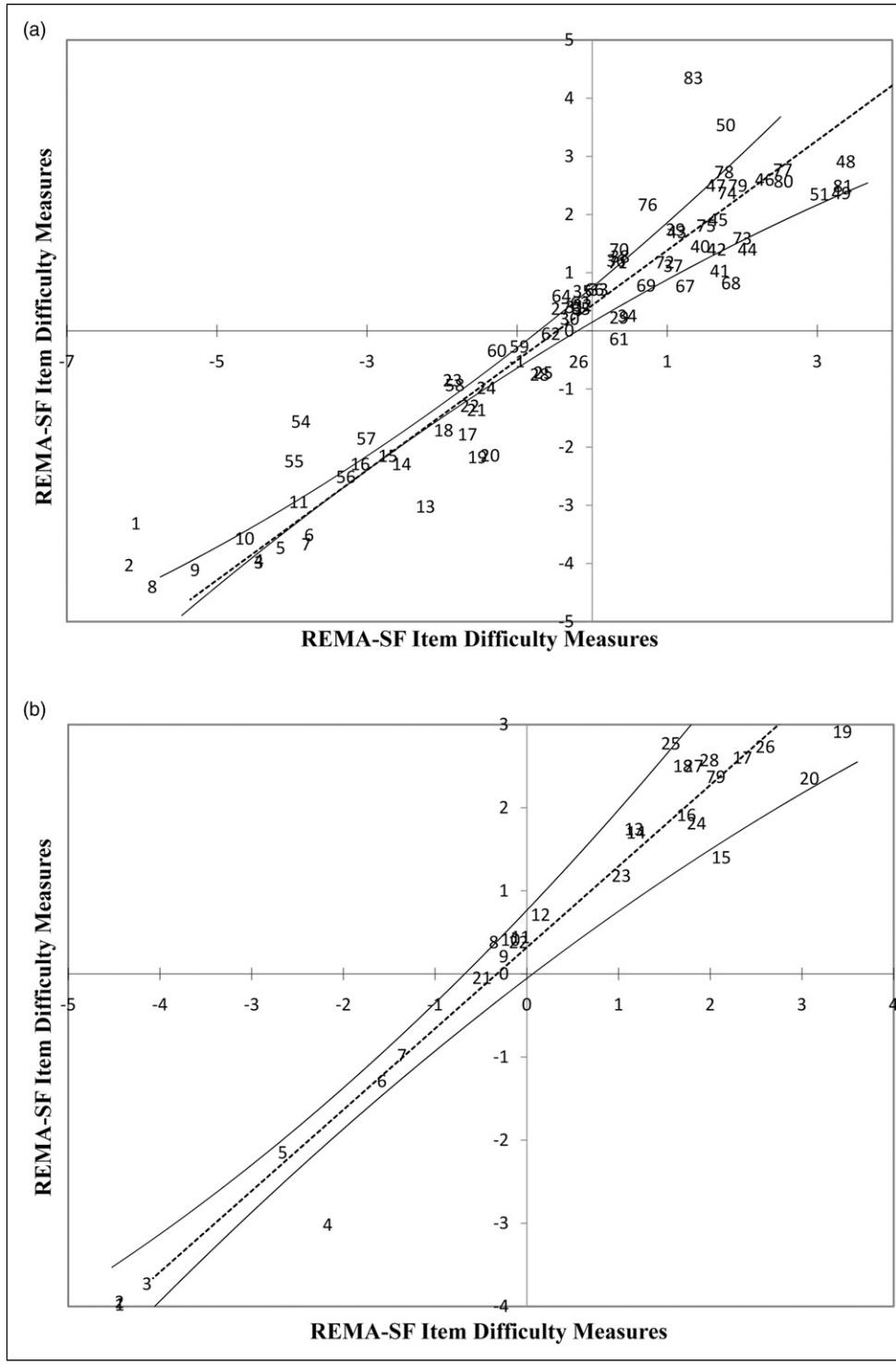

Figure 3. Cross plot for the difficulties of common items between two forms of the REMA. (a) All common items. (b) Refined common items. Notes. Numbers shown in this figure are entry numbers of the correctness indicators.

Table 1. Item Difficulties of Refined Common Items in and T Statistics for Testing the Differences.

		REMA-SF		REMA-Full		
Entry	Item	Measure	S.E.	Measure	S.E.	t-statistic
1	N1_17A	−4.44	.10	−3.97	.05	−1.40
2	N1_17B	−4.44	.10	−3.94	.05	−1.67
3	N1_17C	−4.14	.10	−3.72	.04	−.99
4	N6_N14	−2.17	.06	−3.01	1.01	1.14
5	N1__34	−2.66	.07	−2.14	.04	−2.57*
6	N6__31	−1.58	.07	−1.28	.03	.17
7	N6__44	−1.36	.07	−.97	.04	−.95
8	N7_N27	−.36	.07	.39	.21	−1.97
9	N2__N9	−.25	.09	.22	.04	−1.59
10	N3__49	−.18	.09	.42	.05	−2.79
11	N6__55	−.07	.10	.45	.05	−1.85
12	N5_N12	.15	.07	.72	.17	−1.40
13	N7_N28	1.17	.13	1.75	.21	−1.08
14	N6__60	1.19	.19	1.71	.07	−1.02
15	N6_N32	2.12	.22	1.41	.57	1.67
16	N6__74	1.74	.21	1.92	.14	.53
17	N1__63	2.35	.26	2.61	.10	.19
18	N6__72	1.70	.26	2.51	.14	−1.68
19	N6_N19	3.44	.40	2.92	.39	1.49
20	N6_N35	3.08	.53	2.36	.62	1.27
21	G4__N4	−.49	.07	−.04	.10	−1.12
22	G6__28	−.09	.07	.39	.09	−1.46
23	G4__N8	1.03	.16	1.19	.12	.77
24	G1_N16	1.85	.20	1.82	.35	.85
25	G4_N52	1.57	.21	2.78	.56	−1.50
26	G5__45	2.60	.29	2.74	.15	.53
27	G6_N25	1.82	.27	2.51	.66	−.53
28	G7__37	1.99	.29	2.58	.51	−.47
29	G8_N27	2.06	.19	2.38	.82	−.01

Notes. * indicates significant differences at 0.05 level.

shown in both studies one and two, the short test can distinguish between high and low performers reliably, as does the long version. We also observed adequate overlapping variance between the two forms of the REMA from the score equating results. In addition, the REMA-SF has demonstrated sufficient validity evidence on an independent sample in study two (i.e., cross-observer validity).

More importantly, the REMA-SF can offer substantive time and resource savings, which is a major incremental value of using the short form to assess early mathematics. For example, training assessors and coders on the 80-item REMA-SF costs less and requires less time than the training on the 197 REMA-Full items. We believe the REMA-SF can overcome the challenge of young children's short attention spans in assessment practice while retaining comprehensive coverage and sufficiently reliable psychometric properties.

Can the REMA-SF be Equated with the Widely Used REMA-Full?

In the third study, we equated the REMA scores between two assessment forms, allowing for scores of participants who took different REMA tests to be meaningfully compared. In the literature, a wide range of approaches (e.g., standardized scores, linking, calibration, and equating) was developed to compare scores across tests, and equating has been perceived as the most stringent (Dorans et al., 2010; Yu & Osborn-Popp, 2005).

The Educational Testing Service (ETS) group has suggested five requirements for equating, including equal construct, equal reliability, symmetry, equity, and population invariance (Dorans et al., 2010; Holland & Dorans, 2006). The REMA-SF's construct and reliability are equal with the full version. Symmetry requires the transformation for mapping the scores of the REMA-SF into the REMA-Full to be inversely performed. This can be easily achieved based on the established transformation formula. Moreover, there should be no difference in the form administered to children to measure their math competencies in future practices, which addresses the equity requirement.

Regarding the population invariance requirement, it remains an empirical question about whether the equating function is population invariant regardless of the choice of any sub-population. However, a substantive amount of research has suggested that absolute invariance of a measure or its equating function across any sub-population is unattainable in real-world settings (e.g., Dong & Dumas, 2020; Livingston, 2004). Therefore, we believe the current study still produces valid equating between two forms of the REMA, even though the last requirement has yet to be tested. This issue also points to an interesting and important future direction, that is, whether the developed REMA-SF can measure early mathematical cognition in the same way over different groups (e.g., developmental stages, demographics, and measurement occasions). The measurement invariance of the REMA is a further validity issue to consider in future research.

Implications and Conclusions

The REMA score equating in this study will open new paths for investigating important issues in education. For example, the generalizability of educational intervention effects has been a concerning issue to researchers and educators (Tipton & Olsen, 2018). The REMA-Full was used to measure children's mathematics competencies in large-scale intervention studies (e.g., Clements et al., 2020, 2011, 2013), and the REMA-SF has also been used to collect non-interventional data in different regions of the United States. The score equating procedures may enable early math researchers to compare REMA scores across studies and further investigate the generalizability of intervention effects. In addition, revisions have been made in the REMA assessment across years, so the score equating may help maximize the use of the high-quality early childhood cognitive data that are often not easy to collect.

In conclusion, our study demonstrated that the research-based early mathematics assessment-short form (REMA-SF) is a sound measure of children's math skills in early childhood, and the validity of the measure was well-generalized to an external sample. Additionally, we also equated the REMA scores between the long and short forms of the assessment: anchor items common across the forms were selected and refined in the equating process. Therefore, the REMA-SF is a valid assessment for future research and practice.

Acknowledgment

The authors wish to express appreciation to the original school districts, teachers, and children who participated in this research Portions of this work were supported by the DREME Network based at Stanford University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Funding Sources: Heising-Simons Foundation, National Science Foundation (Grant / Award Number: 'ESI-9730804', 'REC-0228440'), and Institute of Education Sciences (Grant / Award Number: 'R305A110188', 'R305K05157').

ORCID iD

Yixiao Dong https://orcid.org/0000-0003-1940-952X

Supplementary Material

Supplementary material for this article is available online. http://journals.sagepub.com/doi.suppl/10.1177/07342829211037195

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). JAM Press.

Biddlecomb, B., & Carr, M. (2011). A longitudinal study of the development of mathematics strategies and underlying counting schemes. *International Journal of Science and Mathematics Education*, *9*, 1-24. https://doi-org.du.idm.oclc.org/10.1007/s10763-010-9202-y

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Taylor & Francis. https://doi.org/10.1111/j.1745-3984.2003.tb01103.x

Bracken, B. A. (1984–1998). *Bracken basic concept scale-revised*. The Psychological Corporation, Harcourt Brace and Company.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *29*(1), 3-20. https://doi.org/10.2307/749715

Clements, D. H., Dumas, D. G., Dong, Y., Banse, H. W., Sarama, J., & Day-Hess, C. A. (2020). Strategy diversity in early mathematics classrooms. *Contemporary Educational Psychology*, *60*, 101834, https://doi.org/10.1016/j.cedpsych.2019.101834

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*(6045), 968-970. https://doi.org/10.1126/science.1204537

Clements, D. H., & Sarama, J. (2021). *Learning and teaching early math: The learning trajectories approach* (3rd ed.). Routledge. https://www.routledge.com/Learning-and-Teaching-Early-Math-The-Learning-Trajectories-Approach/Clements-Sarama/p/book/9780367521974.

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., & Fesler, L. (2020). Effects on mathematics and executive function of a mathematics and play intervention versus mathematics alone. *Journal for Research in Mathematics Education*, *51*(3), 301-333. https://doi.org/10.5951/jresemtheduc-2019-0069

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, *28*(4), 457-482. https://doi.org/10.1080/01443410701777272

Clements, D. H., Sarama, J., Wolfe, C. B., & Day-Hess, C. A. (2017). REMA-SF --Research-based Early Mathematics Assessment Short Form. Denver, CO: *Kennedy Institute, University of Denver.*

Clements, D. H., Sarama, J., & Day-Hess, C. A. (2021). Review of assessments of early childhood mathematics competencies. *Submitted for publication*.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127-166. https://doi.org/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *American Educational Research Journal*, *50*(4), 812-850. https://doi.org/10.3102/0002831212469270

Dong, Y., & Dumas, D. G. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences*, *160*, 1-23. https://doi.org/10.1016/j.paid.2020.109956

Dong, Y., Fan, W., Cheung, F. M., & Li, M. (2020). Development of a short form of the CPAI-A (Form B) with Rasch analyses. *Journal of Applied Measurement*, *21*(4), 515-532.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating (ETS Research Rep. No. RR-10-29)*. ETS.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan, & R. Murnane (Eds.), *Whither opportunity? Rising inequality and the uncertain life chances of low-income children* (pp. 47-70). Sage.

Fuson, K. C., Smith, S. T., & Cicero, A. M. L. (1997). Supporting latino first grader's ten-structured thinking in urban classrooms. *Journal for Research in Mathematics Education*, *28*(6), 738-760. https://doi.org/10.2307/749640

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Praeger.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program user's guide*. Winsteps.com. http://www.winsteps.com/.

Linacre, J. M. (2021). Winsteps® Rasch measurement computer program. Winsteps.com.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. ETS. https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf.

McCrae, R. R., & Costa, P. T., (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences*, *28*(3), 116-128. https://doi.org/10.1027/1614-0001.28.3.116

OECD. (2014). *Strong performers and successful reformers in education - Lessons from PISA 2012 for the United States*. OECD Publishing. https://doi.org/10.1787/9789264207585-en

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge. https://doi.org/10.4324/9780203883785.

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2012). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *5*(2), 105-135. https://doi.org/10.1080119345747.2011,627980.

Shepard, L. A. (1994). The challenges of assessing young children appropriately. *The Phi Delta Kappan*, *76*(3), 206-212. https://www.jstor.org/stable/20405297

Siegler, R. S. (1993). Adaptive and non-adaptive characteristics of low income children's strategy use. In L. A. Penner, G. M. Batsche, H. M. Knoff, & D. L. Nelson (Eds.), *Contributions of psychology to science and mathematics education* (pp. 341-366). American Psychological Association.

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). JAM Press.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102-111. https://doi.org/10.1037/1040-3590.12.1.102

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524. https://doi.org/10.3102/0013189x18781522.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352-360. https://doi.org/10.3102/0013189X14553660.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-johnson III tests of achievement* (3rd ed.). Riverside.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370-371.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Yu, C. H., & Osborn-Popp, S. E. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment, Research, and Evaluation*, *10*(1), Article 4. https://doi.org/10.7275/68dy-z131