Journal for Research in Mathematics Education

The Journal for Research in Mathematics Education is an official journal of the National Council of Teachers of Mathematics (NCTM). *JRME* is the premier research journal in mathematics education and is devoted to the interests of teachers and researchers at all levels—preschool through college.

ARTICLE TITLE:

A Validity Argument for a Brief Assessment of Mature Number Sense

AUTHOR NAMES:

Kirkland, Patrick K.; Cheng, Ying; McNeil, Nicole M.

DIGITAL OBJECT IDENTIFIER: 10.5951/jresematheduc-2022-0071

VOLUME: 55

ISSUE NUMBER: 1

Mission Statement

The National Council of Teachers of Mathematics advocates for high-quality mathematics teaching and learning for each and every student.

CONTACT: jrme@nctm.org

NCTM | NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS

Journal for Research in Mathematics Education
2024, Vol. 55, No. 1, 51–67

A Validity Argument for a Brief Assessment of Mature Number Sense

Patrick K. Kirkland, Ying Cheng, and Nicole M. McNeil
University of Notre Dame

This Brief Report presents an example of assessment validation using an argument-based approach. The instrument we developed is a Brief Assessment of Students' Mature Number Sense, which measures a central goal in mathematics education. We chose to develop this assessment to provide an efficient way to measure the effect of instructional practices designed to improve students' number sense. Using an argument-based framework, we first identify our proposed interpretations and uses of student scores. We then outline our argument with three claims that provide evidence connecting students' responses on the assessment with its intended uses. Finally, we highlight why using argument-based validation benefits measure developers as well as the broader mathematics education community.

Keywords: Number sense; Argument-based validation; Assessment

Too often, the mathematics education community has relied on unsystematic, informal assessments to evaluate advocated instructional practices. Some researchers have called for a greater focus on validation evidence to increase transparency and trustworthiness (e.g., Bostic et al., 2019; Carney et al., 2022; Krupa et al., 2019), but approaching validation as outlined in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014; henceforth the Standards) remains rare for assessments in mathematics education.

In this Brief Report, we respond to this call by presenting a validity argument for a new Brief Assessment of Mature Number Sense. We use Kane's (2013) argument-based approach to validity in which the main goal is to communicate the specific proposed interpretations and uses of our assessment's scores. Within this approach, tests or even test scores themselves are not "valid" or "invalid." Rather, as outlined in the Standards, validation is a process that requires identifying the purpose of using test scores and connecting evidence to that purpose. The Standards encourage evidence from five potential sources, including test content, response processes, internal structure, relation to other variables, and consequences of testing (see the Standards for more details).

We first describe our construct of mature number sense and present our brief assessment along with the intended uses of scores from the assessment. We then outline our validity argument, specifying "the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses" (Kane, 2013, p. 1). We summarize the evidence collected to date, including evidence from test content, students' response processes, and the test's internal structure, and we communicate validation as the process of evaluating the strength of evidence in support of each claim. Finally, we discuss the advantages of using a validity-argument approach, keeping in mind that validation should be viewed as an ongoing process in which developers continue to improve the strength of their argument.

Mature Number Sense

During the last three decades, "number sense" has emerged as a central goal of mathematics education. In the late 1980s, government-commissioned reports (e.g., National Research Council, 1989) and national curriculum frameworks (e.g., National Council of Teachers of Mathematics [NCTM], 1989) highlighted number sense as a core objective of K–12 mathematics education. Students exhibiting number sense have the disposition to make sense of numerical situations and use a rich conceptual understanding of number and operations to flexibly solve problems (McIntosh et al., 1997). For example, when shown the problem $24 \times 25 = 12 \times \_\_$, someone with number sense may see that 12 is half of 24 and then know they need to double 25 to keep the equation balanced. Someone not exhibiting number sense, yet still displaying procedural understanding, would probably solve the problem multiplying $24 \times 25 = 600$ and then dividing by 12. Likewise, someone exhibiting number sense would see a problem like $26 \times 3/4$ and know automatically that the product will be less than 26 (without needing to multiply).

This research was supported by a Community-Engaged Research Grant from the Center for Social Concerns at the University of Notre Dame and the National Science Foundation under Grant No. DRL EHR 2100214. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Thanks go to Allison VanOverberghe and Joanna Azar for their help with coding and administrative tasks and to Guangjian Zhang for statistical consultation. This research would not have been possible without the support of many administrators and teachers throughout the country, especially Kathe Streeter. The preregistration for this study is available at https://osf.io/xk97w/.

Beyond mathematics education, however, number sense has been operationalized and assessed in a variety of ways across the disciplines of neuroscience, developmental psychology, and special education (see Whitacre et al., 2020, for a review). Therefore, per the recommendation of Whitacre et al. (2020), we include the term "mature" to distinguish our construct from approximate number sense (e.g., Dehaene, 2001) and early numeracy (e.g., Jordan et al., 2009). Importantly, this does not imply that maturity in number sense naturally develops with age or that our construct is distinct from prior work in mathematics education (Markovits & Sowder, 1994). In this project, we define *mature number sense* in line with the definition by McIntosh et al. (1992) of number sense as "a person's general understanding of number and operations along with the ability and inclination to use this understanding in flexible ways to make mathematical judgements" (p. 3).

Mature number sense is a central objective in mathematics education, and many instructional practices have been designed to promote it (e.g., Number Talks, Parrish, 2011; Number Sense Routines, Shumway, 2018). Such practices are popular with mathematics teachers, yet none has been evaluated in the U.S. using a measure of number sense with available validity evidence for student scores. Instead, evaluations have relied on varied sources, ranging from classroom observations (e.g., Parrish, 2011) to standardized tests of general mathematics achievement (e.g., Boaler et al., 2018).

Measuring Mature Number Sense

Several mature number sense measures with validity evidence exist, but they are primarily used outside of the U.S. context (see Table A1). McIntosh et al. (1997) developed the Number Sense Tests, paper-and-pencil tests designed to gather evidence of 8- to 14-year-old students' number sense in many countries in response to curricular frameworks' focus on number sense (e.g., NCTM, 1989). Questions were multiple choice, and students were encouraged to answer in under 30 s. Overall administration time was 30 min. To discourage the use of algorithms, students were told not to write anything but the answer on each page. The number of items ranged from 30 to 45 depending on age, and the items were balanced across six strands of number sense: (a) understanding the meaning and size of numbers, (b) understanding and use of equivalent representations of numbers, (c) understanding the meaning and effect of operations, (d) understanding and use of equivalent expressions, (e) computing and counting strategies, and (f) measurement benchmarks. Student and teacher interviews on the items, as well as reliability statistics, provided validity evidence that scores reflected students' current level of number sense.

Yang et al. (2008) built on this assessment by developing computer-administered tests that included the ability to prompt students for their reasoning and confidence rating for each answer (e.g., Yang, 2019). These additions increased the required administration times to 75–80 min. Scores have evidence of construct validity according to item review by mathematics educators and student interviews as well as a factor analysis confirming the authors' structural hypotheses from studies of 9- to 12-year-old Taiwanese students. For sixth-grade students, the authors hypothesized five components: (a) understanding the meaning of numbers and operations, (b) recognizing number size, (c) using multiple representations of numbers and operations, (d) recognizing the relative effect of operations on numbers, and (e) judging the reasonableness of computational results.

After reviewing these measures, we chose to develop a new Brief Assessment of Mature Number Sense in hopes of providing a practical assessment that could be used to evaluate the impact of number-sense-focused instructional practices and to advance the study of mature number sense as a construct. We reasoned that long administration times may be one reason the existing assessments have not been used to evaluate number-sense-focused instructional practices in the U.S. Could we develop an assessment that performs as well as these instruments while keeping average administration time around 10 min?

Research Question

The question guiding our work was: What evidence exists for the claim that student scores on the Brief Assessment of Mature Number Sense can be interpreted as reflective of students' mature number sense and can be used to detect growth that may result from instruction designed to promote number sense?

A Brief Assessment of Mature Number Sense

We interpreted the McIntosh et al. (1992) definition to include four components of mature number sense: (a) understanding number concepts and magnitude, (b) using multiple representations of numbers, (c) understanding the effect of arithmetic operations on numbers, and (d) understanding mathematical equivalence. Inclusion of these components was based on prior conceptualizations of mature number sense (e.g., McIntosh et al., 1997; NCTM, 1989; Yang, 2019). We did not include what McIntosh et al. (1997) called "measurement benchmarks" (p. 41) and NCTM (1989) called "developing referents for measures of common objects" (p. 40) because these items rely on culturally specific knowledge of measurement units and reflect a magnitude understanding of a measurement unit rather than numbers and operations. We also did not include

"judging the reasonableness of estimates of computed results" from Yang et al. (2008, p. 789) because this component overlaps with the items across multiple components in which students must estimate a result and choose the best answer. We hypothesized, similar to Yang (2019), that mature number sense would be an overarching hierarchical latent construct, with the four components each theoretically related to mature number sense. That is, a student's understanding of the effect of operations on a number reflects their mature number sense as well as a more specific understanding of how operations affect computational results.

Table 1 presents the items of our brief assessment along with their identified component. The answer choices presented to students are included in the last column, with the correct answer bolded.

Interpretation and Use Statement

Following the recommendations set out by Carney et al. (2022), we offer an interpretation and use statement designed to act as the assessment's abstract. Its purpose is to provide an "initial evaluation point for the end user, as well as an explicit statement of the intended interpretations for proposed uses to support development or examination of a validity argument" (Carney et al., 2022, p. 335).

Our Brief Assessment of Mature Number Sense (Table 1) measures individual students' capability to make sense of numerical situations and use a rich conceptual understanding of number and operations to flexibly solve problems (McIntosh et al., 1997). Improving students' number sense is a central goal in mathematics education (e.g., National Governors Association Center for Best Practices [NGO Center] & Council of Chief State School Officers [CCSSO], 2010) and many instructional practices have been designed to promote students' number sense. The assessment is a 23-item, electronic multiple-choice test in which students have 60 s to solve each item mentally. If students do not respond after 60 s, the assessment automatically advances to the next item. The time remaining for each item is not visible to students. The target population is sixth- to 12th-grade students. Students complete the assessment individually using an electronic device (e.g., Chromebook). The assessment can be administered in an individual or group setting and is a freely available educational resource. Median student completion time is under 10 min. If students choose not to answer an item, it is counted as incorrect. Scores are generated as latent trait estimates using the mirt package in R and then converted to a scaled score (range 50–150). Scaled scores represent the current level of a student's mature number sense (in this sample, $M=110$, $SD=20$). However, we have not yet established a valid population norm for scores. The assessment was designed as an efficient way to reliably measure U.S. students' number sense to detect growth that may result from instruction designed to promote number sense. The assessment may be used by researchers or practitioners interested in studying the effects of number-sense-focused instructional practices. Scores are not meant to be interpreted as reflective of comprehensive grade-level standards-based proficiency, nor should they be used in any evaluative purpose with high-stakes implications. Although the brief assessment was informed by international studies, validity evidence has not yet been gathered from international samples. Thus, we recommend its current application be limited to U.S. students.

Validity Argument

In support of the intended interpretations and uses of the scores from our Brief Assessment of Mature Number Sense, evidence was collected from several sources over a series of studies. Here we organize that evidence with respect to each associated claim in our validity argument.

Claim 1: Students' Response Patterns Reflect the Use of Mature Number Sense When Solving an Item Correctly

We gathered evidence from a study on test content (Study 1) and on students' response processes (Study 2) to support this claim.

Study 1: Expert Item Review

To develop a list of items, we first specified our population of interest to be sixth- to 12th-grade students. We chose sixth grade as the starting point because national standards (e.g., NGO Center & CCSSO, 2010) suggest that students would have already been exposed to the content of our potential items. This means that, by the time they take this assessment, students would have been exposed to topics such as percentages, division with fractions, and decimal magnitude. We intentionally included high school students, given our interest in eventually examining how number sense develops during schooling.

We then constructed an item pool from previous measures (McIntosh et al., 1997; Yang et al., 2008), studies of mathematical equivalence (Rittle-Johnson et al., 2011), popular standardized tests (e.g., National Assessment of Educational Progress [NAEP], Trends in International Mathematics and Science Study [TIMSS]), and our own teaching experience. The items were reviewed internally for alignment with our construct definition, grade-level appropriateness of content, redundancy, and clarity. We pilot-tested a 36-item version of the measure with local sixth- to eighth-grade students ($N=254$)

Table 1

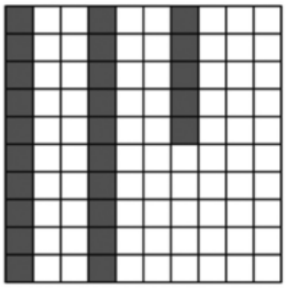
Brief Assessment of Mature Number Sense

Order	Component[a]	Item	Answers (**correct**)
1	MR	Which is true about the number 2/5?	• It is greater than 1/2. • It is the same as 2.5. • **It is equivalent to 0.4** • It is less than 1/6.
2	EQ	24 × 25 = 12 × ___	• 12.5 • 25 • **50** • 100
3	MR	The sale price of milk at a store is 25 percent off the regular price. Which of the following statements means the same thing?	• 1/25 off regular price • 1/5 off regular price • **1/4 off regular price** • 2/5 off regular price
4	NCM	Which of the following numbers could you put in the blank to make this number statement true? 1/2 × ___ = 3/6	• **1** • 2/4 • 3/2 • 3
5	NCM	What number is 8 thousands + 4 hundreds + 5 ones?	• 548 • 845,000 • 845 • **8,405**
6	MR	Which of the following numbers could be represented by the point A on the number line? 	• 0.6 • 2.06 • 2.4 • **2.6**
7	EO	Which is the best estimate for 87 × 0.09?	• **A lot less than 87** • A little less than 87 • A little more than 87 • A lot more than 87
8	EO	Which total is larger than 1?	• 2/5 + 2/7 • **4/7 + 1/2** • 1/2 + 4/9 • 2/8 + 1/11
9	NCM	Use two of the numbers below to make a fraction that is as close as possible to 1/2. 3, 4, 9, 12	Dropdown menus with the four possible choices for both numerator and denominator (**4/9**)
10	EQ	4 × 36 = 4 × (___ + 6)	• 6 • 3 • **30** • 576
11	NCM	Which fraction represents the largest amount?	• **5/6** • 5/7 • 5/8 • 5/9
12	MR	Which is true about the number 5?	• It is equivalent to 0.5 • It is equivalent to 5% • It is equivalent to 5/0 • **It is equivalent to 500%**
13	EQ	7 + 12 + ___ = 12 + 8 + 7	• 19 • **8** • 46 • 712

(*Continued*)

Table 1 (*Continued*)

Brief Assessment of Mature Number Sense

Order	Component[a]	Item	Answers (**correct**)
14	NCM	Which is larger, 4/5 or 11/12?	• 4/5 • **11/12** • They are equal • Can't know
15	EO	Given that 784+496=1280, what is 7.84+4.96?	• .1280 • 1.280 • **12.80** • 1280
16	EQ	100×16−2×16=___×16	• 86 • **98** • 102 • 25,568
17	MR	Each grid is made up of 100 small squares that are all the same size. What part of the grid is shaded? 	• 0.2 • **0.25** • 2 • 2.5
18	EQ	18×5=9×___	• **10** • 36 • 90 • 810
19	EQ	6×24=(6×___)+(6×4)	• 4 • 6 • **20** • 144
20	EO	Select the best estimate for 12 ÷ 1/5	• Less than 12 • Equal to 12 • **Greater than 12** • Can't know
21	EQ	6×(3×2)=(___×3)×2	• 2 • 3 • **6** • 36
22	EO	Select the best answer for 26×3/4	• **Less than 26** • Equal to 26 • Greater than 26 • Can't know
23	NCM	93×134=12,462. Use this to find the answer to: 12,462 ÷ 930.	• 10 • **13.4** • 93 • 134

[a]EO=effect of operations; EQ=equivalence; MR=multiple representations; NCM=number concepts and magnitude

and removed items that did not meet preestablished statistical criteria (e.g., item-total correlations below .3, item discrimination values below .2).

We then surveyed relevant experts to rate items according to their "importance for assessing mature number sense." As outlined in the Standards, "important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA et al., 2014, p. 14). One manner in which developers can collect this evidence is by soliciting expert judgments of test items.

Participants. Six research and three teaching experts reviewed the potential items. The research experts were identified as having PhDs and at least 5 years of experience researching K–8 mathematics education or mathematical concepts such as algebraic reasoning or arithmetic. The teaching experts were identified as meeting at least one of the following qualifications: winner of a teaching excellence award such as the presidential teaching award; National Board Certified; or serving as a mathematics specialist in their school or district.

Methods. Experts were asked to rate items on a scale of 1 to 5 (according to the item's "importance for mature number sense": 1=*not important*, 3=*important but not essential*, 5=*essential*). They were presented the following description at the beginning:

> We are aware that the construct of "number sense" can be operationalized differently across disciplines. In this study, we are interested in what some researchers have called "mature number sense"—this construct encompasses multidigit and rational number sense. It is what is most commonly referred to as number sense in math education. For this survey, we are not interested in the approximate number system or early childhood numeracy (e.g., cardinality, early counting).

Items rated included our target items intermixed with distractor items. We calculated the mean rating for each target item and removed any items with a mean rating below 3.0 ("*important*").

Results. For the 23 items in our brief assessment, the mean expert rating was 3.91 (out of 5.0; $SD=0.42$; $M=2.5$ for intermixed distractor items). None of the retained items received a rating of 1 or 2 by an expert or were marked problematic in the comments.

Study 2: Student Think-Aloud Interview

To gather information about student strategy use, we conducted a retrospective think-aloud protocol. Here, we focus on evidence that links students' response processes with the construct. This is especially important given the multiple-choice and timed nature of our brief assessment. Questions are timed to help differentiate between the use of standard algorithms and flexible, number sense strategies.

Participants. Participants were middle ($n=3$) and high school students ($n=6$) in a local summer school. The summer school program is designed for both remediation and enrichment, depending on the students' needs. Demographic information for the participants is presented in Table A2.

Methods. Students were individually presented items on a Chromebook in a classroom with an experimenter present. After answering, students were prompted to "explain in your own words how you solved that problem." Students' explanations were untimed. To facilitate explanation, the item appeared on the screen along with the student's chosen answer. If something was unclear in a student's verbal response, the experimenter would ask additional follow-up prompts. After the student gave an explanation, the assessment advanced to the next item.

Sessions were audio recorded and transcribed. From there, student responses were coded according to strategy used. For each item, example student responses for each code were generated. Student strategies were coded as one of the following: Arithmetic (use of traditional algorithm); Number Sense (use of a variety of item-specific number sense strategies, such as estimation, rational number knowledge, properties of operations, and equivalence); Incorrect "Number Sense" (use of a strategy demonstrating a misunderstanding, [e.g., "When you multiply, it becomes a greater number"]); Guess and Check; and Unknown (no clear strategy provided). Using the earlier example of $24\times25=12\times\_\_$, the student who said, "I ran out of time, but I tried to multiply them with 24 on top and 25 on bottom" was coded as using Arithmetic. The student who said, "12 is half of 24, so then I did half of 25 to keep it even" was coded as using Incorrect "Number Sense."

One experimenter coded the set of student responses (270 answers), and then a second coder independently coded a randomly selected sample of 20% of the responses (54 answers) for a reliability check. We attained 91% agreement between the coders. After analyzing the disagreements together, the original code on one solution was changed. Coders reanalyzed all codes for that item and agreed that all other original codes were correct. The coded strategies were analyzed for both correct and incorrect responses.

Results. When answering items correctly, students used number sense strategies 77% of the time and arithmetic strategies only 11% of the time. In contrast, when students solved items incorrectly, they demonstrated a misconception 39% of the time (e.g., "I thought when you divided, it would get smaller"), used arithmetic 29% of the time, guessed and checked 14% of the time, and had an unknown strategy 17% of the time. Frequently when students used arithmetic and were incorrect, they reported running out of time (e.g., "I couldn't do the mental math in time").

Claim 2: In the Absence of Number-Sense Focused Instruction, Student Scores on the Brief Assessment of Mature Number Sense are Internally Consistent, Fair, Reliable Over Time, and Reflective of Our Hypothesized Framework

If we intend for student scores to be used to evaluate instructional practices designed to promote number sense, we need to provide evidence that the scores are reliable reflections of students' number sense in the absence of number-sense-focused instruction. In addition, we need to gather evidence that the assessment reflects our hypothesized framework of mature number sense. To do this, we field-tested our brief assessment with students from nine schools whose mathematics teachers had not (yet) incorporated Number Talks or similar number-sense-focused instructional practices. Here we present evidence related to the internal structure of the assessment, or the relationships among items, as well as the reliability of students' overall scores.

Study 3: Field Test Administration

Participants. Participating schools included one public school (N of students = 236) for a one-time administration and eight private schools for a test-retest administration (N of students = 605 for Round 1). Participating students approximated the racial diversity of the U.S. K–12 student population (47% White, 20% Hispanic, 11% Black, 10% two or more races/ethnicities, 6% other races/ethnicities, 5% Asian; 2020 U.S. K–12 Public School Enrollment: 46% White, 28% Hispanic, 15% Black, 5% two or more races/ethnicities, 5% Asian, NCES, 2020). Given this similarity, as well as the diverse nature of the communities served by these schools, we believe our sample, although not a perfect random sample because of the nature of nonpublic schools, generalizes to the broader population of U.S. sixth- to 12th-grade students. Demographics for participating students are presented in Table A2 and for schools in Table A3.

Methods. Students completed the measure in class using school-provided devices. For the test-retest administration, teachers were sent a link to share with students. After their class completed the first round, we sent teachers a second link for students to complete approximately 2 weeks later (M = 15.2 days, SD = 5.65). We use the linked Round 1 and 2 results (N = 483) to examine the test-retest reliability of scores. In all other analyses, we do not include the Round 2 results.

Analysis. To measure internal consistency, we chose to calculate multiple measures of internal consistency reliability to demonstrate that our conclusion does not change on the basis of the choice of measure. We report Cronbach's alpha (Cronbach, 1951) and the Spearman-Brown formula (Brown, 1910; Spearman, 1910) on the even and odd items. In addition, we analyzed each item for the presence of differential item functioning (DIF; Drasgow et al., 2018) by subgroup (e.g., gender, race, native language) using the Wald test. DIF occurs when subgroups of students with similar overall performance have systematically different performance on a single item. To measure test-retest reliability, we used a Pearson correlation of student scores from the two administrations.

Results. On average, students correctly solved 12.95 of the 23 items (56%, SD = 4.83). This is slightly higher than in prior studies using other measures of mature number sense (e.g., 37–50% in McIntosh et al., 1997). The median item response time was 17.1 s (M = 21.56, SD = 15.11), and the median overall completion time was 7.4 min (M = 7.58, SD = 2.6). Detailed demographic-level and item-level performance are available in Tables A4 and A5, respectively.

Internal consistency was high according to both Cronbach's alpha and the Spearman-Brown formula (α = .83; r = .81). The DIF analyses did not reveal any systematic bias in items across gender, race/ethnicity, or native language. No items produced unexpected differences after correcting for a false discovery rate. For the 483 matched participants in the test-retest, the correlation between scores in Round 1 and 2 was r = .84. Because all coefficients reported here are greater than the benchmark r = .7, we have sufficient evidence of scores' consistency internally, as well as over time.

Factor Analyses

The hypothesized structure of the assessment was mature number sense as the primary trait, encompassing four distinct components. On the basis of this structure, one could consider fitting a second-order factor model, in which mature number sense acts as the superordinate factor, with indirect effects through the hypothesized components. In this model, a participant's mature number sense determines their level for a specific component, which in turn determines the observed item

score. An alternative approach to represent a similar latent structure is the bifactor model. Crucially, in this model, the latent factor for mature number sense is directly measured by item performance and is modeled independently of the four component-specific factors (Cai, 2010). A participant's specific component levels are related to their observed scores, but their mature number sense is also directly related to every observed score (Rodriguez et al., 2016). That is, a student correctly answering an "effect of arithmetic operations" item directly reflects their mature number sense as well as a more specific understanding of how operations affect computational results. We felt that this interpretation aligned better with our hypothesized framework and proceeded with testing the bifactor model against plausible alternatives.

Methods. All models were fit using the mirt package in R, with a difficulty or intercept parameter and a discrimination or slope parameter for each item. We tested this against a unidimensional model in which all items loaded directly on a mature number sense factor (Model 2) to see whether we needed to include the components. We also fit a simple structure four-factor model (Model 3) and a correlated four-factor model (Model 4), in which items loaded directly on four component factors. In these models, no latent factor represents number sense. Finally, we tested the second-order factor model (Model 5) described earlier. A path diagram of each model is presented in Figure A1.

We examined model fit using the following fit indices: the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMSR), and comparative fit index (CFI). Values less than 0.05 for RMSEA and SRMSR indicate a good fit. For CFI, values greater than .95 suggest a strong model fit (Rodriguez et al., 2016). Finally, because the models are nested, we compared model fits using a chi-square difference likelihood ratio test.

Results. Detailed model-fit statistics are available in Table A6. The hypothesized bifactor model fit the data well: RMSEA = 0.030; 90% CI for RMSEA [0.026, 0.035]; SRMSR = 0.039; CFI = .981. According to all indicators used, the bifactor model fit better than any competing model, although the unidimensional and second-order models also fit well. Therefore, we then tested these models using a likelihood ratio test to account for additional parameters used across models. The bifactor model fit significantly better than the unidimensional model ($\Delta\chi^2(23) = 114.24$, $p < .001$), as well as the second-order model ($\Delta\chi^2(23) = 1381.74$, $p < .001$). These results provide evidence that student response patterns reflect our hypothesized framework of mature number sense. Detailed item parameter estimates are available in Table A7.

We then used the bifactor model to generate latent trait estimates for students' mature number sense using the mirt package in R. Latent trait estimates are generated for each participant with a multivariate normal prior distribution with a mean of 0. We focus on the latent trait estimate for the overall factor as reflective of students' mature number sense given the evidence presented in Claim 1 and the factor structure evidence presented here. We then use a linear transformation to convert the trait estimates to a more interpretable scale score (range 50–150, $M = 110$, $SD = 20$ in this sample).

Claim 3: Student Scores on the Brief Assessment Are Malleable When Exposed to Number-Sense-Focused Instructional Practices

If we intend for our brief assessment to evaluate instructional practices designed to promote number sense, we must provide evidence that student scores are malleable, not fixed. Here, we present evidence from a pilot study with sixth-grade teachers implementing Number Talks, a number-sense-focused instructional practice (Parrish, 2011).

Study 4: Exposure to Number Talks

Participants. We partnered with a local middle school on a Number Talks and Restorative Justice initiative. The school's two sixth-grade mathematics teachers agreed to consistently implement the intervention in their classrooms during a semester ($M = 1.5$ Number Talks per week), whereas the seventh- and eighth-grade teachers could not commit to doing so ($M = 0.25$ per week). Claim 3 suggests that sixth-grade students' ($n = 68$) scores on our assessment should improve, whereas the seventh- and eighth-grade students' scores ($n = 112$) should not.

Methods. Before the professional development sessions, students completed our brief assessment following the same protocol described earlier. Teachers were then trained on using Number Talks and provided a Number Talks curriculum according to student results. After 4 months, students once again completed the assessment. Student scores were matched and then analyzed.

Results. Consistent with Claim 3, sixth-grade students' number sense scores improved 25% from before the training in January ($M = 8.26$, $SD = 3.74$, $\theta_{6thPre} = 94.2$) to posttest in May ($M = 10.31$, $SD = 4.09$, $\theta_{6thPost} = 99.6$), reflecting significant growth, $t(67) = 3.39$, $p = .001$, whereas seventh- and eighth-grade students' scores did not ($M_{Pre} = 11.69$, $SD = 5.01$, $\theta_{Pre} = 105.7$, $M_{Post} = 11.96$, $SD = 5.00$, $\theta_{Post} = 106.6$, $t(111) = 0.95$, $p = .34$). Although our purpose was not to test effectiveness

of a particular number-sense-focused instructional practice, this finding provides some evidence for malleability because we would expect number sense to be malleable to intentional number-sense-focused instruction.

Discussion

We have presented a validity argument for our Brief Assessment of Mature Number Sense. Users of this assessment can efficiently measure and monitor mature number sense and compare scores before and after instructional practices designed to promote students' number sense.

This work serves as an example of argument-based validation for mathematics education assessments. A major advantage of using a validity-argument approach (Kane, 2013) is that the inferences and assumptions made in linking an observed test score to the interpretations and uses of that test score are made explicit. For example, skeptics may argue that a multiple-choice test cannot truly measure a students' number sense, but we address that here by clearly outlining evidence in support of our inference that these items assess number sense (Claim 1).

Overall, the level of evidence in support of our first two claims is strong. Students demonstrate evidence of using number sense strategies when they solve the items correctly, the internal consistency of scores is high, and items are fair to all participants (e.g., no DIF; Claim 2). For students without number-sense-focused instruction, scores were stable, whereas scores improved for students who received number-sense-focused instruction in the form of Number Talks (Claims 2 and 3).

Currently, we have limited information on how scores relate to other variables, such as students' grade-level mathematics achievement and understanding of other mathematics constructs. We also do not yet know how mature number sense develops as students advance from middle to high school or how scores on the measure correlate with scores from previously developed, time-intensive measures of number sense (e.g., Yang, 2019). Finally, we should note that designing a brief assessment required us to make choices about the types of items and overall weight of each component, and other researchers may not agree with all our choices. We aimed to produce a practically useful measure of the construct, but it is not meant to be the only measure.

We can use our brief assessment to address these limitations by continuing the ongoing validation process. We can examine how scores correlate with those on other relevant measures, such as the longer Yang (2019) measure, grade-level mathematics achievement tests and tests of domain-general cognitive functioning. Future studies with a broader sample of schools can gather evidence to establish a valid population norm for mature number sense. Additionally, we intend to expand the brief assessment to include students in third to fifth grades, vertically scaling two forms of the assessment, so researchers can study both the development of number sense across time and the impact of instructional routines such as Number Talks in elementary grades in which they were first used (Humphreys & Parker, 2015).

By sharing the details of our validity argument, our goal is to demonstrate the benefits of this approach for both instrument developers and the broader field. We had to make certain decisions about the length of the test, types of items, and weighting across our components given our primary goal of developing a brief assessment. Other developers may choose to focus their assessments differently. However, by openly sharing our process and our items, our hope is that the field can build on this validity evidence to grow our collective understanding of students' mature number sense.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Boaler, J., Dieckmann, J. A., Pérez-Núñez, G., Sun, K. L., & Williams, C. (2018). Changing students' minds and achievement in mathematics: The impact of a free online student course. *Frontiers in Education*, *3*, Article 26. https://doi.org/10.3389/feduc.2018.00026

Bostic, J. D., Krupa, E. E., Carney, M. B., & Shih, J. C. (2019). Reflecting on the past and looking ahead at opportunities in quantitative measurement of K–12 students' content knowledge. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 205–229). Routledge. https://doi.org/10.4324/9780429486197-9

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 1904–1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Carney, M. B., Bostic, J., Krupa, E., & Shih, J. (2022). Interpretation and use statements for instruments in mathematics education. *Journal for Research in Mathematics Education*, *53*(4), 334–340. https://doi.org/10.5951/jresematheduc-2020-0087

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dehaene, S. (2001). Précis of the number sense. *Mind & Language*, *16*(1), 16–36. https://doi.org/10.1111/1468-0017.00154

Drasgow, F., Nye, C. D., Stark, S., & Chernyshenko, O. S. (2018). Differential item and test functioning. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 885–899). Wiley. https://doi.org/10.1002/9781118489772.ch27

Humphreys, C., & Parker, R. (2015). *Making number talks matter: Developing mathematical practices and deepening understanding, Grades 3–10*. Stenhouse.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867. https://doi.org/10.1037/a0014939

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Krupa, E. E., Carney, M., & Bostic, J. (2019). Argument-based validation in practice: Examples from mathematics education. *Applied Measurement in Education*, *32*(1), 1–9. https://doi.org/10.1080/08957347.2018.1544139

Markovits, Z., & Sowder, J. (1994). Developing number sense: An intervention study in Grade 7. *Journal for Research in Mathematics Education*, *25*(1), 4–29. https://doi.org/10.2307/749290

McIntosh, A., Reys, B. J., & Reys, R. E. (1992). A proposed framework for examining basic number sense. *For the Learning of Mathematics*, *12*(3), 2–8. https://flm-journal.org/Articles/94F594EF72C03412F1760031075F2.pdf

McIntosh, A., Reys, B., Reys, R., Bana, J., & Farrell, B. (1997). *Number sense in school mathematics: Student performance in four countries*. Mathematics, Science & Technology Education Centre, Edith Cowan University. https://ro.ecu.edu.au/ecuworks/6819

National Center for Education Statistics. (2020). *Racial/ethnic enrollment in public schools*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/cge/racial-ethnic-enrollment

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. http://www.corestandards.org.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. National Academies Press. https://doi.org/10.17226/1199

Parrish, S. D. (2011). Number talks build numerical reasoning. *Teaching Children Mathematics*, *18*(3), 198–206. https://doi.org/10.5951/teacchilmath.18.3.0198

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology*, *103*(1), 85–104. https://doi.org/10.1037/a0021334

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Shumway, J. F. (2018). *Number sense routines: Building mathematical understanding every day in Grades 3–5*. Stenhouse.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Whitacre, I., Henning, B., & Atabaş, S. (2020). Disentangling the research literature on *number sense*: Three constructs, one name. *Review of Educational Research*, *90*(1), 95–134. https://doi.org/10.3102/0034654319899706

Yang, D.-C. (2019). Development of a three-tier number sense test for fifth-grade students. *Educational Studies in Mathematics*, *101*(3), 405–424. https://doi.org/10.1007/s10649-018-9874-8

Yang, D.-C., Li, M., & Lin, C.-I. (2008). A study of the performance of 5th graders in number sense and its relationship to achievement in mathematics. *International Journal of Science and Mathematics Education*, *6*(4), 789–807. https://doi.org/10.1007/s10763-007-9100-0

Authors

Patrick K. Kirkland, Institute for Educational Initiatives, University of Notre Dame, Notre Dame, IN 46556; pkirklan@nd.edu

Ying Cheng, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556; ycheng@nd.edu

Nicole M. McNeil, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556; nmcneil@nd.edu

Submitted November 30, 2022

Accepted March 8, 2023

doi:10.5951/jresematheduc-2022-0071

APPENDIX

Supplemental Materials

Table A1

Popular Measures of Mature Number Sense

	Number Sense Test	Three-Tier Number Sense Test	Brief Assessment of Mature Number Sense
Source	McIntosh et al. (1997)	Yang (2019)	Current Brief Report
Description	First large-scale multiple-choice measure of mature number sense, used to gather evidence on current level of number sense of students in the US, Australia, Sweden, and Taiwan.	Rigorously developed multiple-choice measure of mature number sense, in which students are prompted to choose the reason for their answer to an item as well as rate how confident they are in their answer and reason for each item.	Brief multiple-choice assessment rigorously developed for practical use by researchers and teachers in classrooms.
Grade levels	2–9	5 (other versions address 3–6)	6–12
Number of items	30 items for age 8 to 45 items for age 14	40 items with reasons and confidence levels	23 items
Time to administer	~30 min	Two 40-min sessions	~10 min
Strengths	Large international samples, theoretically aligned with a Number Sense framework, large item bank.	Rigorous psychometric evidence, scoring includes student reasoning.	Argument-based validation, brief administration time, operationalized equivalence in line with current research.
Weaknesses	Paper-and-pencil format, unstandardized time limit for items, length of administration, operationalizing equivalence.	Length of administration, differing components by grade, scores for reasons include credit for misconceptions.	Does not include students' reasoning or strategy use, limited evidence at this time on relationship with other variables.

Table A2

Self-Reported Student Demographics

Variable	Study 2	Study 3 Round 1	Study 3 test-retest matched	Study 4 sixth-grade pre-post matched
		Grade level		
5	–	15 (1.8%)	15 (3.1%)	–
6	2 (22.2%)	82 (9.8%)	69 (14.3%)	68 (100.0%)
7	1 (11.1%)	207 (24.6%)	50 (10.4%)	–
8	–	161 (19.1%)	58 (12.0%)	–
9	5 (55.6%)	162 (19.3%)	130 (26.9%)	–
10	1 (11.1%)	113 (13.4%)	92 (19.0%)	–
11	–	56 (6.7%)	43 (8.9%)	–
12	–	45 (5.4%)	26 (5.4%)	–
		Gender		
Male	5 (55.6%)	369 (43.9%)	192 (39.8%)	32 (47.1%)
Female	4 (44.4%)	439 (52.2%)	279 (57.8%)	34 (50.0%)
Prefer not to say	–	33 (3.9%)	12 (2.5%)	2 (2.9%)
		Race/ethnicity		
American Indian or Alaskan Native	1 (11.1%)	16 (1.9%)	8 (1.7%)	3 (4.4%)
Asian	–	40 (4.8%)	30 (6.2%)	–
Black or African American	3 (33.3%)	93 (11.1%)	19 (3.9%)	16 (23.5%)
Hispanic or Latino	1 (11.1%)	166 (19.7%)	114 (22.6%)	5 (7.4%)
Native Hawaiian or other Pacific Islander	–	1 (0.1%)	1 (0.2%)	1 (1.5%)
White	4 (44.4%)	392 (46.6%)	245 (50.7%)	25 (36.8%)
Two or more races/ethnicities	–	86 (10.2%)	42 (8.7%)	14 (20.6%)
Other	–	47 (5.6%)	24 (5.0%)	4 (5.9%)
		Primary home language		
English	8 (88.9%)	682 (81.1%)	381 (78.9%)	55 (80.9%)
Spanish	1 (11.1%)	75 (8.9%)	48 (9.9%)	3 (4.4%)
Other	–	25 (3.0%)	18 (3.7%)	3 (4.4%)
Multiple languages	–	59 (7.0%)	36 (7.5%)	7 (10.3%)
Total	9	841	483	68

Table A3

Publicly Available School Demographics for Study 3

Demographic	School 1	School 2	School 3	School 4	School 5	School 6	School 7	School 8
State	OK	NJ	TN	WI	TX	IL	IL	TX
School type	Private, religious	Private, religious	Private, religious	Private, independent	Private, religious	Private, religious	Private, religious	Private, religious
Grade levels	9–12	7–12	7–12	PK–12	6–12	PK–8	PK–8	9–12
Annual tuition	$500–2,500	$39,900	$9,000 before work study	$18,100 for HS	$10,395 for HS	$5,575	$5,200	$22,900
Additional notes	Work study financial aid	All male	Work study financial aid		All female	Tax credit scholarships available	Tax credit scholarships available	All female
				Enrollment % by race/ethnicity				
American Indian or Alaskan Native	0.9%	–	–	0.3%	0.7%	–	1.4%	0.6%
Asian	2.6%	9.2%	–	9.3%	3.3%	0.4%	10.8%	7.0%
Black or African American	10.4%	2.0%	64.0%	4.7%	2.7%	7.7%	0.7%	2.7%
Hispanic or Latino	69.6%	5.3%	24.4%	6.0%	65.7%	18.3%	42.6%	19.9%
White	13.9%	75.3%	4.7%	70.2%	20.3%	68.3%	41.9%	66.5%
Native Hawaiian or Pacific Islander	–	0.2%	–	–	–	0.8%	–	–
Two or more races	2.6%	8.0%	6.9%	5.0%	7.3%	4.5%	4.1%	3.4%

Table A4

Detailed Performance in Study 3 Round 1

Variable	*N*	Mean correct (%)	*SD*
	Grade level		
5	15	8.9 (38.7%)	4.0
6	82	10.8 (47.0%)	3.6
7	207	10.6 (46.1%)	4.4
8	161	13.2 (57.4%)	4.9
9	162	13.6 (59.1%)	4.2
10	113	15.0 (65.2%)	4.4
11	56	16.4 (71.3%)	4.8
12	45	16.3 (70.9%)	4.0
	Gender		
Male	369	13.7 (59.6%)	5.2
Female	439	12.4 (53.9%)	4.4
Prefer not to say	33	12.5 (54.3%)	4.8
Total	841	13.0 (56.5%)	4.8

Table A5

Item Statistics From Field Test Administrations

#	Item	Mean solved	Median RT (in s)	Mean correct (*p*)	Item disc. (D*)	Item-total correlation (*r*)
18	What is true about the number 2/5?	94%	15.97	0.54	0.77	.54
6	24 × 25 = 12 × ___	89%	23.64	0.58	0.71	.44
48	The sale price of milk at a store is 25 percent off the regular price. Which of the following statements means the same thing?	96%	17.14	0.65	0.79	.56
17	Which of the following numbers could you put in the blank to make this number statement true? 1/2 × ___ = 3/6	96%	19.66	0.23	0.62	.44
49	What number is 8 thousands + 4 hundreds + 5 ones?	98%	12.30	0.89	0.25	.24
43	Which of the following numbers could be represented by the point A on the number line?	98%	12.69	0.87	0.33	.33
11	What is the best estimate for 87 × 0.09?	97%	14.62	0.44	0.73	.48
30	Which total is larger than 1?	94%	20.91	0.47	0.72	.42
9	Use two of the numbers below to make a fraction that is as close as possible to 1/2. 3, 4, 9, 12	90%	29.29	0.36	0.29	.12
3	4 × 36 = 4 × (___ + 6)	93%	13.84	0.69	0.62	.41
21	Which fraction represents the largest amount?	96%	9.65	0.74	0.67	.48
45	What is true about the number 5?	97%	11.95	0.34	0.76	.48
32	7 + 12 + ___ = 12 + 8 + 7	96%	9.32	0.84	0.53	.45
26	Which is larger, 4/5 or 11/12?	96%	11.28	0.43	0.62	.36
22	Given that 784 + 496 = 1280, what is 7.84 + 4.96?	93%	16.73	0.71	0.59	.41
8	100 × 16 − 2 × 16 = ___ × 16	87%	24.73	0.27	0.43	.23
46	Each grid is made up of 100 small squares that are all the same size. What part of the grid is shaded?	97%	21.58	0.70	0.52	.31
38	18 × 5 = 9 × ___	95%	14.56	0.67	0.74	.48
2	6 × 24 = (6 × ___) + (6 × 4)	90%	19.37	0.38	0.40	.20
35	Circle the best answer for 12 divided by 1/5 (one-fifth)	94%	11.45	0.38	0.61	.33
33	6 × (3 × 2) = (___ × 3) × 2	94%	12.58	0.73	0.63	.42
36	Circle the best answer for 12 times 3/4	93%	9.82	0.47	0.65	.43
51	93 × 134 is equal to 12,462. Use this to write the answer to the following: 12462/930	93%	15.72	0.56	0.59	.31

Note. Mean solved is the percentage of time a participant chose an answer for an item, whether that answer was correct or incorrect. For example, participants could skip an item without choosing an answer or they could run out of time on an item before the choosing an answer. The item discrimination (D*) values reported here are different than those reported as part of the IRT model. These are commonly referred to as "index of discrimination" values in the classical test theory literature. For each item, D* is the difference between the proportion in the upper quintile of total scores who answered the item correctly and the proportion in the lowest quintile who answered the item correctly.

Table A6

Results of Model Comparisons: Model Fit Statistics

Model	RMSEA, $\hat{\varepsilon}$	90% CI for RMSEA	SRMSR	CFI	AIC
1. Bifactor	0.030	[0.026, 0.035]	0.039	.981	23,483
2. Unidimensional	0.044	[0.040, 0.047]	0.048	.957	23,631
3. Four-factor: simple structure	0.096	[0.092, 0.099]	0.177	.794	25,051
4. Four-factor: correlated	0.097	[0.093, 0.100]	0.180	.793	24,003
5. Second order	0.040	[0.036, 0.044]	0.071	.951	–

Table A7

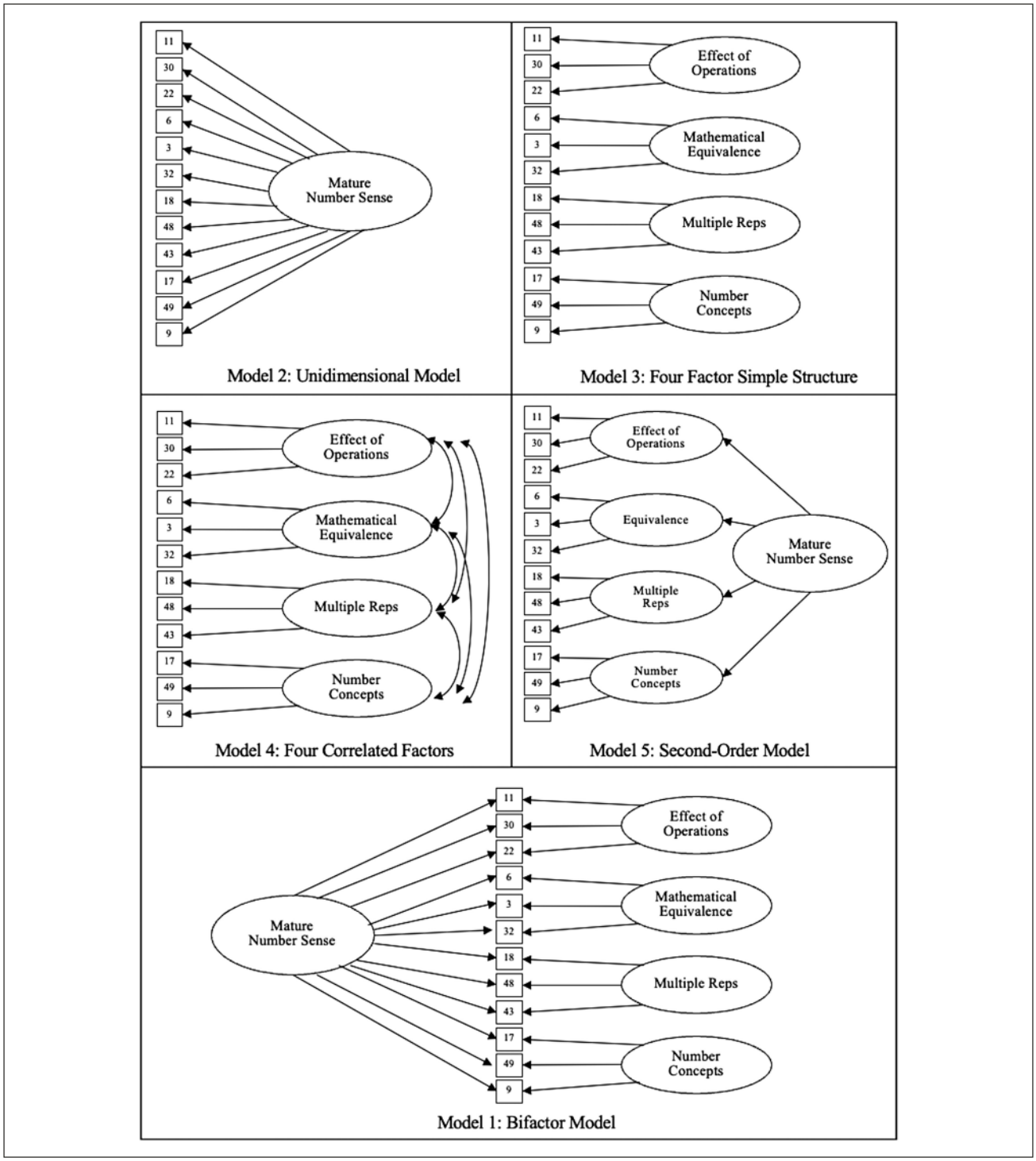
Item Parameter Estimates for Bifactor IRT Model

#	α_{NS}	α_{Oper}	α_{Equiv}	$\alpha_{MultRep}$	α_{NumCon}	*d*
18	1.89	–	–	−0.21	–	0.30
6	1.17	–	0.26	–	–	0.43
48	2.30	–	–	0.08	–	1.20
17	1.43	–	–	–	0.50	−1.64
49	1.01	–	–	–	−0.51	2.51
43	1.53	–	–	−0.004	–	2.61
11	1.40	0.36	–	–	–	−0.29
30	1.20	−0.11	–	–	–	−0.13
9	0.28	–	–	–	−0.21	−0.58
3	1.16	–	0.36	–	–	1.04
21	2.54	–	–	–	−1.17	2.16
45	1.50	–	–	0.09	–	−0.89
32	3.14	–	2.34	–	–	4.28
26	1.99	–	–	–	2.80	−0.64
22	1.29	−0.08	–	–	–	1.21
8	0.58	–	0.30	–	–	−1.08
46	1.75	–	–	2.58	–	1.93
38	1.64	–	1.09	–	–	1.16
2	0.46	–	0.20	–	–	−0.50
35	0.85	−0.06	–	–	–	−0.57
33	1.28	–	0.73	–	–	1.40
36	1.88	2.09	–	–	–	−0.23
51	0.74	–	–	–	0.07	0.28

Note. The α_{NS} column represents the discrimination on the general mature number sense factor. The following columns represent discrimination values on the specific factors: α_{Oper} for effect of operations, α_{Equiv} for equivalence, $\alpha_{MultiRep}$ for multiple representations of numbers, and α_{NumCon} for number concepts and magnitude. *d* represents the multidimensional item difficulty parameter or the item's "easiness." Here, we use the convention from the mirt package in R that higher values reflect easier items and lower values reflect a more difficult item.

Figure A1

Path Diagrams for Each of the Models Tested



Note. For readability, only three items per factor or 12 items in total are represented. Each box represents an observed item score and each oval represents a modeled latent trait or factor. In Model 1, Mature Number Sense is the general factor, whereas the other four factors are all orthogonal specific factors.