Assessing formal knowledge of math equivalence among algebra and pre-algebra students

Emily R. Fyfe, Indiana University

Percival G. Matthews, University of Wisconsin-Madison

Eric Amsel, Weber State University

Katherine L. McEldoon, Tennessee Department of Education

Nicole M. McNeil, University of Notre Dame

This represents the accepted version of the manuscript to be published in

Journal of Educational Psychology

word count (primary text and abstract) = 9106 + 157

Author Note

Emily R. Fyfe, Department of Psychological and Brain Sciences, Indiana University. Percival G. Matthews, Department of Educational Psychology, University of Wisconsin-Madison. Eric Amsel, Department of Psychology, Weber State University. Katherine L. McEldoon, Tennessee Department of Education. Nicole M. McNeil, Department of Psychology, University of Notre Dame. The authors thank Daniel Bolt for helpful guidance on the methodological aspects of the manuscript.

Address correspondence to Emily R. Fyfe, Department of Psychological and Brain Sciences, Indiana University1101 E. 10th Street, Bloomington IN 47405. Email: efyfe@indiana.edu.

Abstract

A central understanding in mathematics is knowledge of *math equivalence*, the relation indicating that two quantities are equal and interchangeable. Decades of research have documented elementary-school (ages 7-11) children's (mis)understanding of math equivalence, and recent work has developed a construct map and comprehensive assessments of this understanding. The goal of the current research was to extend this work by assessing whether the construct map of math equivalence knowledge was applicable to middle school students and to document differences in formal math equivalence knowledge between students in pre-algebra and algebra. We also examined whether knowledge of math equivalence was related to students' reasoning about an algebraic expression. In the study, 229 middle school students (ages 12-16) completed two forms of the math equivalence assessment. The results suggested that the construct map and associated assessments were appropriate for charting middle school students' knowledge and provided additional empirical support for the link between understanding of math equivalence and formal algebraic reasoning.

Keywords: algebra, mathematical equivalence, measurement development, mathematics

Educational Impact and Implications Statement

One of the bedrocks of algebraic thinking is formal knowledge of math equivalence, which is the idea that two sides of an equation are equal and interchangeable. In the present study, we sought to validate a measure of this knowledge in middle school students. Students in pre-algebra and algebra classes were successful on some items, but still struggled with others (such as explicitly defining the equal sign or reasoning about operations on both sides of the equal sign [e.g.,

explaining why 89 + 44 = 87 + 46 is true]). We also found that performance on the measure was related to students' formal understanding of algebraic expressions. Our results highlight the importance of measuring formal knowledge of math equivalence beyond elementary school, particularly given its critical link to algebra.

Assessing formal knowledge of math equivalence among algebra and pre-algebra students

One of the bedrocks of early algebraic thinking is knowledge of *math equivalence*, the relation indicating that two quantities are equal and interchangeable (e.g., Baroody & Ginsburg, 1983; Carpenter et al. 2003; Kieran, 1981; MacGregor & Stacey, 1997). Unfortunately, much research has documented students' (mis)understanding of math equivalence in symbolic form (e.g., Behr, Erlwanger, & Nichols, 1980; Knuth et al., 2005; Li et al., 2008; Lindvall & Ibarra, 1980; Powell & Fuchs, 2010; Weaver, 1973). This research underscores the need for assessments that can both track students' formal knowledge of math equivalence and serve as valid and reliable outcome measures for intervention work. Recent research has developed comprehensive assessments of this understanding among elementary school children, ages 7-11 (Matthews, Rittle-Johnson, McEldoon, & Taylor, 2012; Rittle-Johnson, Matthews, Taylor & McEldoon, 2011). The primary goal of the current study was to investigate whether the assessments could reliably and validly measure formal knowledge of math equivalence among an older cohort of students, ages 12-16, and to report differences in math equivalence knowledge between students in pre-algebra and algebra classes. We also examined how performance on the equivalence measure was related to formal reasoning about a specific algebraic expression.

Knowledge of Math Equivalence

Math equivalence is a broad construct and a formal understanding of it encompasses a number of related components (e.g., Falkner et al., 1999; Charles, 2005; Kieran, 1981; McNeil, 2014). One component is a relational understanding of the equal sign (i.e., knowing that values on either side of the equal sign need to be the same amount). However, other key components include correctly identifying the sides of an equation, noticing relations within equations, encoding equations in their entirety (e.g., noticing the location and order of operators, quantities,

and the equal sign), generating a strategy for equalizing the two sides of an equation (e.g., solving for an unknown), and knowing that a quantity can be represented in many equal and interchangeable ways (e.g., knowing that 7 can be represented as $3 + 4$, $8 - 1$, $2 + 1 + 4$, 7×1, $14/2$, etc.). Of course, children can have an informal understanding of math equivalence without mapping that understanding to the formal symbols (e.g., Mix, 1999; Sherman & Bisanz, 2009). In the current study, we were specifically interested in student's formal understanding of math equivalence – though we use the term "math equivalence" for conciseness.

Math equivalence is considered a "Big Idea" in mathematics because it is both critical to learning mathematics and because it facilitates an understanding of mathematics as a coherent whole, rather than as a set of unrelated concepts and procedures (Charles, 2005; National Council of Teachers of Mathematics [NCTM], 2000). Further, math equivalence is considered a critical pre-requisite to formal algebraic understanding (e.g., Jacobs, Franke, Carpenter, Levi, & Battey, 2007; Kieran, 1981; Knuth, Stephens, McNeil, & Alibali, 2006; MacGregor & Stacey, 1997). Accordingly, the Common Core State Standards recognize the importance of math equivalence and prescribe that children as early as first grade should be able to understand the relational meaning of the equal sign, to determine unknown numbers in equations (e.g., $12 = 5 +$ __), and to judge equations as true or false (e.g., $6 = 6$, $5 + 2 = 2 + 5$; National Governors Association Center for Best Practices and the Council of Chief State School Officers, 2010).

Unfortunately, decades of research in psychology and mathematics education indicate that many elementary-school children (ages 7-11) in the U.S. struggle to understand math equivalence in symbolic form (Behr et al., 1980; Falkner, Levi, & Carpenter, 1999; Fyfe, McNeil, & Borjas, 2015; Fyfe, Rittle-Johnson, & DeCaro, 2012; McNeil & Alibali, 2005; Perry, 1991; Powell & Fuchs, 2010; Renwick, 1932; Rittle-Johnson, 2006; Rittle-Johnson & Alibali,

1999; Seo & Ginsburg, 2003; Weaver, 1973). The broad issue is one of a "cognitive gap"

between arithmetic and algebra (Herscovics & Linchevski, 1994; van Amerom, 2003). Students'

experiences with arithmetic often lead them to view equations operationally as computational

processes to be carried out, rather than structurally as arguments whose truths can be evaluated

by the products of those processes (e.g., Kieran, 1981; Linchevski & Herscovics, 1996; Sfard &

Linchevski, 1994). For example, children often view a problem like $3 + 4 = 5 + \_\_$ as a signal to

carry out a procedure rather than as two related sides or "objects," $(3 + 4)$ and $(5 + \_\_)$, whose

substitutability is expressed by the equal sign. This "gap" leads to specific errors on a variety of

problems assessing knowledge of math equivalence. For example, children often provide an

operational definition of the equal sign – inferring that it means "get the answer" or "the total"

(e.g., Baroody & Ginsburg, 1983; Behr et al., 1980; McNeil & Alibali, 2005). When solving

problems with operations on both sides of the equal sign (e.g., $3 + 4 = 5 + \_\_$), children often fall

into entrenched patterns of calculation and add up all the numbers (12) to write the total in the

blank (e.g., Alibali, 1999; Falkner et al., 1999; Li et al., 2008). Children also tend to reject

problems that are not in standard operations-equals-answer format, claiming that problems such

as $8 = 5 + 3$ are backwards or false (e.g., Behr et al., 1980; Falkner et al., 1999; Li et al., 2008;

Molina & Ambrose, 2006; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003).

The Construct Map for Math Equivalence

Recent work (e.g. Matthews et al., 2012; Rittle-Johnson et al., 2011) has made strides in

elucidating math equivalence knowledge as a construct and creating assessments that pool items

and tasks from previous research in both psychology and math education (e.g., Baroody &

Ginsburg, 1983; Behr et al., 1980; Carpenter, Franke & Levi, 2003; Jacobs et al, 2007; Li et al.,

2008; McNeil & Alibali, 2005; Perry, 1991; Rittle-Johnson & Alibali, 1999; Sherman & Bisanz,

2009; Sternberg et al., 1991; Weaver, 1973). This type of work is critical in order to document the relative difficulties of all the different types of items used in the literature to measure the same construct and to better understand the order in which children's knowledge is constructed. The construct map put forth by Rittle-Johnson et al. (2011) is shown in Table 1. It was derived after an extensive review of the literature on children's math equivalence knowledge.

The construct map contains four levels of increasing sophistication of knowledge. Although the map breaks it into levels to aid in visualization, the underlying knowledge is conceived of as continuous. The levels differ primarily in terms of the types of equation structures with which children are successful. At Level 1, children succeed with the traditional operations-equals-answer structure (e.g., $8 + 4 = \_\_$) and view the equal sign as an operator signal to calculate an answer. At Level 2, children succeed on a wider variety of equation structures, including problems with operations on the right side of the equal sign (e.g., $\_\_ = 8 + 4$) and problems with no operations (e.g., $3 = 3$). At Level 3, children succeed on problems with operations on both sides of the equal sign and recognize a relational view of the equal sign as valid. Finally, at Level 4, children succeed on problems regardless of structure and generate a flexible, relational view of the equal sign. A hallmark of Level 4 understanding is recognizing transformations that maintain the equality of an equation (e.g., adding the same number to both sides of the equal sign) without engaging in full computation strategies (e.g., Alibali et al., 2007; Carpenter et al., 2003; Jacobs et al., 2007; Steinberg et al., 1991).

In two previous studies, researchers documented the construction and validation of comprehensive assessments intended to measure math equivalence knowledge in terms of this construct map (Matthews et al., 2012; Rittle-Johnson et al., 2011). The assessment items varied in type (e.g., equation-solving, equation-structure, equal-sign) and in structural arrangement

(e.g., operations-equals-answer, operations on both sides). The researchers used Rasch modeling procedures to measure item difficulties on a continuous scale. The result was two forms of an assessment that were reliable and valid. Further, the order of the empirically-derived item difficulties supported the hypothesized construct map in two samples. Thus, this measurement work emerged from and was well aligned with the larger literature on math equivalence, and it resulted in a psychometrically-sound assessment tool for elementary school children (ages 7-11).

The Current Study

The goal of the present study was to extend this work to an older cohort of middle school students. Below we outline our three specific aims and the motivation for each.

Our first aim was to investigate whether the construct map and associated assessments could reliably measure knowledge of math equivalence among an older cohort of students (ages 12-16). Previous research indicates that difficulties with math equivalence persist well past elementary school (e.g., Alibali et al., 2007; Jones, Inglis, Gilmore, & Dowens, 2012; Knuth et al., 2006; Li et al., 2008; McNeil et al., 2006; Renwick, 1932). For example, Knuth et al. (2005) found that nearly half of the middle school students in the study provided an operational definition of the equal sign. Similarly, Booth and Davenport (2013) found the average percent correct on a measure of equal sign understanding was close to 50% for a sample of middle school students. Alibali et al. (2007) also demonstrated that fewer than 60% of students at the end of eighth grade provided a relational definition of the equal sign. More importantly, middle school students, high school students and adults sometimes exhibit operational views of equations even after they are taught about the equal sign or equivalence more generally (e.g., Herscovics & Kieran, 1980; Sfard & Linchevski, 1994). For example, educated adults still sometimes solve standard equivalence problems (e.g., $6 + 8 + 4 = 7 +$ __) using operational

strategies, giving the answers 18 or 25 – particularly under speeded conditions (e.g., Chesney et al., 2013; McNeil & Alibali, 2005; McNeil et al., 2010).

This body of research suggests that valid measures of math equivalence knowledge that function beyond elementary school are clearly needed. However, no studies to date have used a discriminating assessment based on a construct map to unpack the structure of math equivalence knowledge in students beyond age 11. We sought to fill this gap by validating the construct map and associated assessments in an older sample of middle school students.

Our second aim was to report quantitative and qualitative differences in math equivalence knowledge between students in pre-algebra and algebra classes. Many studies have considered middle school students as a homogenous population and have not focused on differences as a function of experience. However, a formal understanding of math equivalence is widely regarded as a necessary component of success in algebra (e.g., Carpenter et al., 2003; Jacobs et al., 2007; Knuth et al., 2006; NCTM, 2000; Steinberg et al., 1991), and students in formal algebra courses are more likely to have experiences that explicitly attempt to bridge the "cognitive gap" between arithmetic and algebra. That is, they are more likely to have experiences that rely on explicit recognition of the arithmetic properties of algebra (e.g., performing the same operation on both sides of the equation). A key question is whether these experiences in algebra support shifts in students' understanding of equivalence. We expect they do and that students in algebra classes will provide more relational responses on the equivalence assessment than pre-algebra students.

Our third aim was to examine the relation between knowledge of math equivalence and students' formal reasoning about algebraic expressions. Prior work suggests that knowledge of math equivalence is related to performance on algebraic equations with variables (e.g., Alibali et al., 2007; Booth & Davenport, 2013; Knuth et al., 2005). For example, middle school students

who provided a relational definition of the equal sign were almost twice as likely to solve algebra equations correctly (e.g., $4m + 10 = 70$) than students who did not (Knuth et al., 2006). But, if math equivalence is truly foundational for algebraic thinking, it should predict performance on a variety of tasks – including tasks with algebraic expressions that do not include the equal sign.

Again, the notion of a "cognitive gap" between arithmetic and algebra is relevant (Linchevski & Herscovics, 1996; van Ameron, 2002). As noted, experiences with arithmetic can give rise to misconceptions about the equal sign and equivalence more generally (e.g., the idea that the equal sign is unidirectional and operational). In a similar way, experiences with arithmetic can also give rise to misconceptions about variables and expressions with variables (e.g., Kinzel, 1999; Lucariello et al., 2014; MacGregor & Stacey, 1997; McNeil, Weinberg, et al., 2010). For example, consider a 12-year-olds' difficulty assigning meaning to expressions such as $3a$, $a + 3$, and $3a + 5a$ because, "there is no equal sign with a number after it" (Kieran, 1981, p. 324), or consider a 13-year-olds' difficulty subtracting $8x$ from both sides of an equation because "I don't know how much is $8x$" (Sfard & Linchevski, 1994, p. 107). Both cases exemplify an operational viewpoint in which students treat expressions in terms of processes – signals to fill in a missing value – rather than objects that can be manipulated. Thus, in broader terms, students' concept of equation and all that it encompasses progresses from an operational view to a relational or structural view (Kieran, 1981; Sfard & Linchevski, 1994), and depending on where students are in this development, we would expect similar kinds of errors in thinking across both the math equivalence assessment *and* their reasoning about algebraic expressions.

Method

Participants

Participants represented a convenience sample of 229 students (106 female) from two public schools in a small city in the American West. One school served grades 7 through 9 ($n =$ 165) and had a student population that was 50% Caucasian, 48% Hispanic, and 86% qualified for free or reduced-price lunch. The other school ($n = 64$) served grades 7 through 12 and had a student population that was 81% Caucasian, 13% Hispanic, 3% African-American, and 29% qualified for free or reduced-price lunch. All students from six different teachers were invited to participate with no exclusion criteria. The majority of participants in this study (94%) were seventh- and eighth-grade students between 12 and 14 years old (*M age* = 13.1 years, *SD* = 0.8, *min* = 12.0, *max* = 16.0). Because one school spanned grades 7-12, some students in the sample were in higher grades and somewhat older than the typical middle school sample. The majority of students (62%) were in pre-algebra classes. The remaining students were in Algebra I (34%) or an advanced secondary mathematics course (3%) that included a mix of algebra and geometry. Because our focus was on class experience (i.e., pre-algebra vs. algebra) rather than grade level, we refer to all students in our sample as "middle school" students. For conciseness, we also refer to both Algebra I students and advanced secondary math students as "algebra" students as these courses included algebra instruction. No records were collected regarding whether students required an individualize education plan (IEP) or testing accommodations.

Materials

Math Equivalence Assessment. We administered two forms of the assessment from Matthews et al. (2012; see also Rittle-Johnson et al., 2011) with a few changes detailed below. Note that Rittle-Johnson et al. (2011) originally constructed two different, but comparable forms in order to facilitate pre-post testing in intervention studies. We continued using two comparable forms in the current design instead of settling on one for similar practical purposes.

There were three problem types. *Open-equation-solving* items assessed students' abilities to solve equations of varying difficulty. For example, an easy item included an operation on the right side of the equal sign (e.g., $8 = 6 + \_\_$), and more difficult items included operations on the left and/or on both sides of the equal sign (e.g., $7 + 6 + 4 = 7 + \_\_$). Several open-equation-solving items also included letter variables (e.g., $c + c + 4 = 16$). *Equation-structure* items assessed students' understanding of valid equation structures as well as their abilities to reason about both sides of an equation without computation. For example, one easier item had students evaluate whether $31 + 16 = 16 + 31$ was true or false, and a more difficult item had students explain why $67 + 86 = 68 + 85$ was true without adding the numbers together. *Equal-sign* items assessed students' explicit understanding of the symbolic equal sign. A core item asked students to define the equal sign, and another item had students rate different definitions of the equal sign as good or not good. A full list of the items used on both forms is included in the Appendix.

Each form of the assessment had 31 items (12 open-equation-solving, 13 equation-structure, and 6 equal-sign). Each item was classified *a priori* as tapping knowledge at Levels 1, 2, 3, or 4 of the construct map (see Table 1). We began with the forms from Matthews et al. (2012), which used a step-by-step item matching procedure to ensure similarity of content and difficulty across forms. However, we introduced a change to ensure that all parameters could be easily placed on the same measurement scale. One form of the assessment we used was identical to Form 1 from Matthews et al. (2012). The other was nearly identical to Form 2, with the key exception that we replaced a number of Form 2 items with those from Form 1 so that we would have 10 anchor items (denoted with the superscript A in the Appendix and in Table 4). We chose anchor items to span the range of item difficulties specified *a priori* by the construct map and distributed across different item types. We oversampled items hypothesized to be more difficult

because we expected pre-algebra and algebra students to have higher ability than past samples, which only included elementary school children. The result was that 10 of the 31 items on each assessment were identical across forms, serving as anchors to aid in equating scores across forms. We applied a concurrent calibration procedure (Kolen & Brennan, 2004) to yield item and person parameters that were on a common metric. Due to experimenter error, one item on Form 1 was misprinted. It was originally intended to be an open-equation-solving item at Level 3 with operations on both sides of the equal sign. However, it was misprinted as __ = 8 + 5 + 9. Since this only contains operations on the right side of the equal sign, the misprint led us to designate it as a Level 2 item rather than a Level 3 item. This is marked in the Appendix (Form 1, Item 25).

Each item was scored dichotomously as correct (1) or incorrect (0). As in prior work, equation-solving answers within one of the correct answer were scored as correct to account for minor calculation errors (e.g., Perry, 1991; Rittle-Johnson, 2006). Ten items required students to provide a written definition or explanation, and responses were coded based on the system from Matthews et al. (2012). Specifically, responses were coded as correct if the student mentioned the equality relation between values on the two sides of the equal sign. For example, for defining the equal sign, responses of "it means the same as" or "the same amount" were coded as correct and responses of "the answer" or "the sum" were not. A second rater independently scored responses for 35% of the sample, and inter-rater agreement was high, with an average agreement of .95 on both Form 1 (range = .89 – 1.00) and Form 2 (range = .88 – 1.00).

Algebraic Expression. In addition to the equivalence assessments, we administered an item that has been used to assess conceptual understanding of algebraic expressions (see McNeil, Weinberg, et al., 2010, see also Küchemann, 1978): "Cakes cost c dollars each and brownies cost b dollars each. Suppose I buy 4 cakes and 3 brownies. What does $4c + 3b$ stand for?" For

students who completed Form 1 of the assessment, the specific symbols used were mnemonic in

that the price of a cake was represented by *c* and the price of a brownie was represented by *b*. For

students who completed Form 2 of the assessment, the specific symbols used were traditional in

that the *c* and *b* were replaced by the traditional letters *x* and *y*. This contrast has been of interest

to researchers because use of mnemonic symbols may strengthen students' naïve conceptions

that variables in algebraic expressions stand for labels instead of quantities (McNeil et al, 2010;

see also Küchemann, 1978, MacGregor & Stacey, 1997; Rosnick, 1981).

This problem differed in key ways from the items on the assessment that contained letter

variables (i.e., $10 = z + 6$, $c + c + 4 = 16$, $m + m + m = m + 12$ on Form 1). First, the problems on

the assessment contained the equal sign, but the algebraic expression did not. Second, the

problems on the assessment required students to solve for the value of the variable, but the

algebraic expression required students to conceptually interpret the symbols without any

calculation. Third, none of the problems on the assessment featured products symbolized by the

juxtaposition of a variable and a coefficient (e.g., $3x$), which is a more difficult symbolic form to

understand relative to a stand-alone variable. Thus, this problem tapped students' formal

understanding of symbolic letter variables and their interpretation within an algebraic expression.

Responses were coded based on a system developed in previous research (McNeil,

Weinberg, et al., 2010). A response was scored as correct (1) if the student indicated that the

letters stood for the cost or price of the cakes and brownies (see Table 2 for examples). A second

rater scored responses for 35% of the sample, and inter-rater agreement was high for coding

responses as correct or incorrect (.95).

Procedure

Assessments were administered to whole classes using an alternating procedure (i.e., we alternated handing out Form 1 and Form 2 so the first student got Form 1, the second student got Form 2, the third student got Form 1, the fourth student got Form 2, etc.). This ensured that about the same number of each form was distributed in each class ($n_{V1} = 114$, $n_{V2} = 115$) and that the distribution would result in randomly equivalent groups taking the two forms (Kolen & Brennan, 2004). The algebraic expression problem was printed as the final problem on a separate page of the assessment. The sessions lasted about 45 minutes. The groups assigned to different test forms were similar in terms of mean age (Form 1 = 13.0, Form 2 = 13.1), percent female (Form 1 = 46%, Form 2 = 47%), and percent in pre-algebra (Form 1 = 63%, Form 2 = 62%).

Data Analysis

We used a Rasch model to examine performance on the assessment. Rasch modeling is a one-parameter member of the item response theory (IRT) family (Bond & Fox, 2007). The Rasch model estimates item difficulty and student ability levels simultaneously, yielding the probability that a particular respondent will answer a particular item correctly (Rasch, 1993; Wright, 1977). We used Winsteps software (3.80.1; Linacre, 2013) to perform all IRT estimation procedures using default settings. Given our use of a common-item design, we applied a concurrent calibration procedure to ensure that the parameter estimates for each form were calibrated to the same scale (e.g., Kolen & Brennan, 2004). Because this approach assumes the common items function equivalently across groups (i.e., invariance), we conducted a check on this assumption by performing separate calibrations of each form and examining the relationship between difficulty estimates for common items across forms. The presence of invariance would be supported by a strong linear relationship between estimates. Specifically, we inspected a scatterplot and found the best fitting linear estimate comparing the estimates from Form 1 and

Form 2. As can be seen in Figure 1, the best-fitting line had a slope of 1.02, an intercept of .05, and all items were close to the regression line. This supports the invariance assumption and suggests that concurrent estimation of all items from both forms is warranted (Kolen & Brennan, 2004; Linacre, 2016). Thus, all estimates discussed below are from the concurrent estimation (of 53 total items, if anchors are not double counted) and can be interpreted against a common scale.

Results

First, we examine the psychometric properties of the math equivalence assessment to determine whether it functions well for students in middle school. Second, we compare performance as a function of students' current math course (pre-algebra vs. algebra). Finally, we describe performance on the algebraic expression item, and test whether students' knowledge of math equivalence is correlated with their interpretation of the algebraic expression.

Math Equivalence Assessment

Rasch model fit information supported the unidimensionality of the assessment, indicating that it largely tapped a single construct. Unidimensionality in Rasch modeling is often assessed by principal components analysis (PCA, e.g., Bond & Fox, 2007; Hattie, 1985; Smith, 1996). Specifically, the model accounted for 45.6% of the variance in our data set (eigenvalue of 43.6). The largest secondary factor accounted for 2.6% of the variance (eigenvalue of 2.5). The Rasch model can also be evaluated using *infit* and *outfit* statistics, which indicate different types of problematic items (Bond & Fox, 2007). Infit statistics detect unexpected responses to items with difficulty estimates close to respondents' ability estimates. In contrast, outfit statistics tend to reflect the influence of unexpected responses to items that are far from participant's ability level. All items on the assessment had good infit values within the range of 0.5 and 1.5 (Linacre, 2016; Wright & Linacre, 1994). By contrast, 19 items had outfit statistics outside the .5 to 1.5

range, and 17 of those items were easy items with near-ceiling performance ($M_{accuracy} = 96.2\%$,

$SD = 2.6\%$). These outfit results were perhaps to be expected given that the assessment was

originally intended for elementary-school aged children and, as such, included several lower-

level items that 7th-9th grade students should complete easily. We followed the advice of Bond

and Fox (2007) and Linacre (2016) and focused more on infit than on outfit measures. See Table

3 for item difficulty estimates and fit measures.

The PCA results along with infit indices support the use of Rasch analysis. As a

supplementary analysis, we also conducted two sets of confirmatory factor analysis (CFAs)

using MPlus software (Muthén & Muthén, 1998-2017) for each assessment version: (a) a 1-

factor model that included all items as loading on a single factor, and (b) a 3-factor model that

separated items according to each of the three question types (structure, equal sign, and solve),

while allowing each of the three factors to co-vary with the others. The results are summarized in

Table 4. For both versions of the assessment, RMSEA for the 1-factor model was ≤.067 and

essentially equivalent to the RMSEA for the 3-factor model (as indicated by nearly identical

confidence intervals). Moreover, the ratio of chi-square to degrees of freedom was relatively low

for the 1-factor model. Although the CFI for both the 1- and 3-factor models was low compared

to a desired benchmark of about .9, this was probably due in part to the lower level of statistical

dependence observed across items and in part due to the large number of items on each form.

The low CFI can be interpreted as a sign that the single underlying dimension is not particularly

strong, although that does not preclude the unidimensionality assumption being appropriate

(indeed, CFI is not high for the 3-factor model either).

On balance, our analyses suggest that the unidimensionality assumption proved adequate

with the caveat that a number of items demonstrate some degree of misfit – primarily indicated

by outfit measures – due to items with low difficulty levels. This issue seems to be an inherent

difficulty of using an assessment that includes measures for the lowest levels of our construct

map with middle school students (e.g., solve _ + 5 = 9). Generally speaking, items with

difficulties that are very low for the sample in question simply yield low information and can be

prone to some sort of misfit. We briefly return to this issue when addressing limitations of our

method in the discussion section.

Beyond demonstrating adequate evidence for the unidimensionality assumption, the

assessments were consistent and showed adequate capacity to resolve person ability and item

difficulty estimates. Item reliability as assessed by the Rasch model was generally good ($R_I =$

.98), indicting that sample size was large enough to estimate item difficulty well (Linacre, 2016).

Person reliability was ($R_P = .78$) was just short of the normative cutoff of .80, indicating that

while adequate, more items may be needed to adequately distinguish between high and low

performers. Given that the measures were designed using a younger sample, it is noteworthy that

the item and person reliabilities remained adequate with the current sample.

We also evaluated whether our hypothesized levels of difficulty matched the empirical

estimates. Recall that we selected items to tap four different levels of knowledge, as outlined in

the construct map (Table 1). The hypothesized level of difficulty for each item (1, 2, 3, or 4; see

Appendix) correlated highly with the empirically-derived item difficulty estimates, Spearman's

$\rho(52) = .88, p < .001$. We further used a Wright map (Wilson, 2005) to visually inspect the

difficulty of the items (see Figure 2). A Wright map has two columns, one for respondents and

one for items. Respondents with higher ability estimates are near the top of the map and those

with lower ability estimates are near the bottom. Similarly, items with higher difficulty estimates

are near the top of the map and items with lower difficulty estimates are near the bottom. The

vertical line indicates the scale for the ability and difficulty estimates measured in logits (i.e.,

log-odds unit). Average difficulty was set to 0 logits. We added horizontal lines to visually

highlight the clustering of items by hypothesized level. However, it should be explicitly noted

that the construct is a continuous measure and that the link between student ability estimates and

item difficulty estimates is a probabilistic one. Thus, the lines we added for clarity should not be

interpreted as discrete stages.

As shown on the Wright map and in Table 3, the items we had *a priori* categorized as

Level 4 items proved to be the most difficult (i.e., clustered near the top of the Wright map). All

of the Level 4 items had higher difficulty estimates than any of the Level 1, 2, or 3 items. The

items we *a priori* categorized as Levels 1, 2, and 3 were somewhat less distinct, but tended to

cluster in the expected order. Most Level 3 items had difficulty estimates near -1, Level 2 items

had difficulty estimates near -2, and Level 1 items had difficult estimates near -2.5.

There was a small set of items (9 out of 62, 14.5%) that did not function according to our

hypothesized levels. Five were open-equation-solving items that were easier than expected. Two

Level 2 items ($8 = 6 + \_\_$, $7 = \_\_ + 3$) functioned more like Level 1 items. This occurred in prior

work as well (Matthews et al., 2012; Rittle-Johnson et al., 2011), suggesting we need to re-

evaluate where these items fall on the construct map. Three Level 3 items ($3 + 6 = 8 + \_\_$, $\_\_ + 2$

$= 6 + 4$, $5 + \_\_ = 6 + 2$) functioned more like Level 2 items, which could be due to error/noise or

to these items functioning differently among this older, more experienced sample.

There were four items that were harder than expected. Three were Level 1 items (pick a

pair equal to $6 + 3$ and judge $8 = 5 + 10$ as true/false on both forms). However, these items'

difficulty estimates were rank ordered similarly to that reported in previous research (Matthews

et al., 2012; Rittle-Johnson et al., 2011). The last was a Level 2 item, the misprinted solve item

on Form 1 (__ = 8 + 5 + 9), perhaps more difficult because it contained three addends instead of two. However, only four students missed this item suggesting it was still relatively easy.

Overall, the assessment performed well, and the construct map seemed to apply to this older sample. Specifically, different item types with varying levels of difficulty were measured on a single scale and functioned in a way that matched the hypothesized construct map. We acknowledge a caveat to this point: The sample studied here is different from the sample in earlier studies (Matthews et al., 2012; Rittle-Johnson et al., 2011), so we cannot directly compare item difficulty estimates. Thus, to evaluate similarity of the construct map's performance and applicability across studies with different age groups, we relied on the similarity of the rank orderings (i.e., rankings from low difficulty to high difficulty estimates) across samples.

Math Equivalence Performance as a Function of Math Course

Students performed well on the assessment (M_{acc} = 79% [25 out of 31], SD = 14%), but only 2% of the sample scored at ceiling. An analysis based on percent correct revealed differences as a function of math course after controlling for age. Specifically, pre-algebra students scored significantly lower (M = 77%, SE = 1%) than algebra students (M = 83%, SE = 2%), $F(1, 226) = 4.45$, $p = .04$, $\eta_p^2 = .02$. Analyses based on Rasch ability estimates supported these results. As shown on the Wright map, there was an approximately normal distribution of ability estimates. Further, ability estimates (M = 2.57, SD = 1.50, *range* = –1.56 to 6.66) were positively correlated with students' self-reported expected grade in the class (A, B, C, D, or F), Spearman's $\rho(229) = .41$, $p < .001$. Ability estimates also differed significantly by math course after controlling for age, $F(1, 226) = 4.95$, $p = .03$, $\eta_p^2 = .02$, such that pre-algebra students had lower ability estimates (M = 2.3, SE = 0.1) than algebra students (M = 2.9, SE = 0.2).

To further examine this group difference, we looked at the probability of success on items of different difficulty levels for students at different ability levels. The model allows us to calculate the probability of any participant's success on any given item from log-odd units by using the following equation:

$$Pr(success) = \frac{1}{1 + e^{-(\theta-d)}},$$

in which θ is a participant's ability estimate and d is the item difficulty estimate. First, we selected six items from Table 3: two items with the lowest and highest difficulty estimates and four items with difficulty estimates that represented each of the four levels on the construct map (e.g., items we *a priori* categorized as Level 1 had difficulty estimates near -2.5, so one item we selected had an empirically-derived difficulty estimate of -2.52). Second, we calculated the mean ability estimates for the pre-algebra group and the algebra group. Table 5 presents the probability of success for items at different difficulty levels for students at the mean ability level for each group. Importantly, the model predicts substantial differences in performance on typical Level 4 items as a function of math course, but predicts few differences for lower level items, as the typical student in both groups is expected to have high probabilities of success.

We qualitatively explored performance on three Level 4 items on which the differences between pre-algebra students and algebra students were particularly pronounced. The first item was a structure item: "17 + 12 = 29 is true. Without adding the 8, can you tell if 17 + 12 + 8 = 29 + 8 is true or false? How do you know?" Only 39% of pre-algebra students responded correctly compared to 64% of algebra students. A common correct response was to write, "You added the same amount to both sides so it's still equal." Students' incorrect responses revealed key differences. Of the pre-algebra students who answered incorrectly, 47% selected false or don't know (as opposed to true) indicating a conceptual misunderstanding. Their false/don't know

selections were accompanied by explanations underscoring the fact that they thought adding the

eights made the problem unequal (e.g., "if you add 8, it won't equal the same," "you're adding 8

so the answer will go up"). In contrast, of the algebra students who answered incorrectly, only

30% were wrong because they selected false or don't know. The other 70% indicated that the

equation was still true, but they had difficulty explaining their selection without calculating each

side (e.g., "both sides equal 37"). Thus, not only were algebra students more likely to solve the

problem correctly than pre-algebra students, even their errors were more indicative of an

emerging relational understanding of equivalence.

The second item was an equal-sign item: "What does the equal sign mean in the

statement: 1 quarter = 25 pennies?" Sixty-nine percent of pre-algebra students defined the equal

sign relationally compared to 86% of algebra students. Common relational responses were to

write, "the same as" or "the same amount of money." The majority of non-relational responses

did not necessarily reflect misconceptions, but were insufficient to convey that equality was

understood rather than simply parroted back (e.g., "equal" "equal to" "they are equal").

However, some non-relational responses reflected a common, operational misconception of the

equal sign, often prevalent among elementary school students (e.g., "it means the answer to the

problem" "the total"). Indeed, of all the pre-algebra students' incorrect responses, 17% of them

were operational, compared to only 6% of algebra students' incorrect responses.

The third item with pronounced differences between groups was an open-equation-

solving item: "Solve for c in the following equation, $c + c + 4 = 16$." Fifty-nine percent of pre-

algebra students solved this item correctly compared to 73% of algebra students. Most often,

students did not show work and simply wrote 6 as their answer. However, some correct

responses were accompanied by written work and indicated a formal algebraic strategy (e.g.,

combining like terms to get $2c + 4 = 16$, subtracting 4 on both sides, and dividing both sides by 2). Twelve percent of pre-algebra students' correct responses were solved using this algebraic strategy compared to 22% of algebra students' correct responses. Incorrect responses were varied, but fell into one of four categories: blank/don't know (33% of incorrect responses), answering 4 (30%), answering 12 (20%), or other (17%). The proportion of each type of incorrect response was similar for pre-algebra and algebra students.

Overall, these results indicate that pre-algebra and algebra students did well on the math equivalence assessment. However, as expected, pre-algebra students had lower ability estimates than algebra students, and the Level 4 items were particularly key for capturing differences.

Interpreting Algebraic Expressions

Overall, 52% of the students interpreted the algebraic expression correctly by indicating that the variables stood for the costs of the cakes and brownies. We used logistic regression to examine whether the likelihood of interpreting the expression correctly depended on the type of variable used (*x*-and-*y* vs. *c*-and-*b*) and current math course (pre-algebra vs. algebra). For students in pre-algebra, those in the *x*-and-*y* condition were significantly more likely to interpret the expression correctly than those in the *c*-and-*b* condition (59% vs. 36%), $B = 0.94$, $SE = 0.34$, $p = .006$, $OR = 2.56$. For students in algebra, those in the *x*-and-*y* condition and those in the *c*-and-*b* condition were equally likely to interpret the expression correctly (55% vs. 66%), $B = -0.51$, $SE = 0.45$, $p = .25$, $OR = 0.60$. This difference was reflected by a significant variable type by math course interaction, $B = 1.45$, $SE = 0.56$, $p = .01$, $OR = 4.27$. Thus, for pre-algebra students, the use of mnemonic letters interfered with their ability to conceptually interpret the expression, whereas algebra students exhibited a deeper conceptual understanding of the expression that was not influenced by the specific letter variables used.

We also descriptively examined students' errors. The most common error was to use the letters as labels for the objects (e.g., three cakes and four brownies) rather than as the cost of the objects, accounting for 41% of all errors. Other errors included writing a literal translation of the expression (e.g., "3 multiplied by x plus 4 multiplied by y;" 23% of errors), responding in vague, uninterpretable ways (e.g., "it is the equation;" 20% of errors), adding unlike terms (e.g., "it must be 7xy;" 9% of errors), or stating an inability to solve the problem (e.g., "I don't know;" 6% of errors). The letters as labels error was more common for students in the *c*-and-*b* condition (46% of errors) than for students in the *x*-and-*y* condition (35% of errors).

Finally, we tested whether students' knowledge of math equivalence was predictive of their interpretation of the algebraic expression. We used logistic regression to examine whether the likelihood of interpreting the expression correctly depended on empirically-derived estimates of student ability on the math equivalence assessment. We included ability estimates as the primary predictor, as well as math course, assessment form, and students' age as control variables. Ability estimates were significantly predictive of success interpreting the algebraic expression, $B = 0.31$, $SE = 0.10$, $p = .001$, $OR = 1.37$. The remaining predictors were not significant when controlling for the others: math course, $B = 0.53$, $SE = 0.38$, $p = .16$, assessment form, $B = 0.39$, $SE = 0.28$, $p = .16$, and age, $B = -0.07$, $SE = 0.21$, $p = .75$. These results were consistent with our hypothesis that knowledge of math equivalence would be related to students' conceptual understanding of algebraic variables.

Recall that several items on the math equivalence assessment contained literal variables (e.g., "Solve for *c* in $c + c + 4 = 16$."). To ensure that the association between math equivalence knowledge and interpretation of the algebraic expression did *not* depend on the items involving literal variables, we ran a secondary analysis. Specifically, we obtained empirically-derived

ability estimates on the math equivalence assessment items that remained after excluding the five

items with variables (out of 52). Ability estimates were still significantly predictive of success

interpreting the algebraic expression, $B = 0.30$, $SE = 0.10$, $p = .002$, $OR = 1.35$, and the

remaining predictors were not significant when controlling for the others: math course, $B = 0.54$,

$SE = 0.38$, $p = .16$, assessment form, $B = 0.34$, $SE = 0.28$, $p = .22$, and age, $B = –0.07$, $SE = 0.21$,

$p = .75$. Thus, students' knowledge of math equivalence – even when assessed without items

using literal variables – was related to their conceptual interpretation of algebraic notation.

Discussion

Math equivalence is considered a "Big Idea" in mathematics as it lays a foundation for

algebraic reasoning and for understanding math more generally (Charles, 2005; Jacobs et al.,

2007; Kieran, 1981; Knuth et al., 2006; MacGregor & Stacey, 1997; NCTM, 2000). Thus,

measuring knowledge of math equivalence is of clear importance. The current study extended the

construct-modeling approach to measuring symbolic equivalence knowledge in three ways. First,

we demonstrated that the equivalence assessment and construct map applied beyond elementary

school, performing well with an older cohort of algebra and pre-algebra students. Second, we

documented differences in math equivalence knowledge between students in pre-algebra and

algebra classes, which were primarily captured by the difficult Level 4 items. Third, we

confirmed that students' understanding of math equivalence was related to their interpretation of

an algebraic expression, even after controlling for age and math course and after excluding the

items containing variables on the math equivalence assessment. Below we outline the theoretical,

practical, and methodological contributions of this research as well as potential future directions.

The results of the current study verified the validity of the math equivalence construct

map explicated by Rittle-Johnson and colleagues (2011; see also Matthews et al., 2012) within

an older and more mathematically experienced sample of middle school students. The items on the math equivalence assessment functioned according to the construct map, with key factors predicting item difficulty as hypothesized. This suggests that the difficulty of the equivalence construct has a stable order, supporting a key assumption of the Rasch model (Rasch, 1993; Wright, 1977). Further, this is a good indication of a generalizable assessment that can be used vertically, at least from early elementary school to middle school algebra. This has important practical applications because previous work demonstrates that middle school students continue to struggle with understanding math equivalence (Alibali et al., 2007; Jones, Inglis, Gilmore, & Dowens, 2012; Knuth et al., 2006; Li et al., 2008; McNeil et al., 2006; Renwick, 1932) thus highlighting the need for assessments that can both track students' formal knowledge of math equivalence and serve as valid and reliable outcomes measures for intervention work.

 As in prior work, the ordering of the item difficulties on the math equivalence assessment confirms that the structure of an equation is a key indicator of complexity and is therefore likely to influence performance (e.g., Baroody & Ginsburg, 1983; Matthews et al., 2012; Powell, Kearns, & Driver, 2016; Rittle-Johnson et al., 2011; Weaver, 1973). Specifically, the greater the structure deviates from the standard operations-equals-answer structure, the more difficult the problem is likely to be. This is true regardless of the specific task. For example, Figure 2 shows that open-equation-solving items are not inherently more difficult than equal-sign-definition problems (or vice versa). Rather, the difficulty depends on the structure of the equation and the extent to which the required solution strategy demands engaging arithmetic principles of equivalence as opposed to simple calculation. This has potential practical implications for designing interventions focused on varying problem structures, rather than varying problem tasks per se. Indeed, this is consistent with intervention research that has facilitated understanding of

math equivalence by including practice with non-standard equation structures (e.g., 17 = 9 + 8;

McNeil, Fyfe, & Dunwiddie, 2015; McNeil, Fyfe, Petersen, Dunwiddie, & Brletic-Shipley,

2011) or instruction on the meaning of the equal sign in the context of non-standard equation

structures (e.g., Fyfe & Rittle-Johnson, 2016; Fyfe, DeCaro, & Rittle-Johnson, 2014; Matthews

& Rittle-Johnson, 2009; Perry, 1991; Powell & Fuchs, 2010).

In addition to validating the construct map, the assessment had considerable resolving

power to detect variability in student knowledge. Even though students did well overall, there

were reliable knowledge differences between students in pre-algebra and algebra. In particular,

the model predicted substantial differences in performance on typical Level 4 items. Students in

algebra were in fact more likely to exhibit comparative relational understanding by reasoning

about transformations that preserve equality without reverting to calculation (e.g., "if we know

17 + 12 = 29, can we tell if 17 + 12 + 8 = 29 + 8 is true without adding?", see Alibali et al., 2007;

Matthews et al., 2012; Steinberg et al., 1991). These Level 4 items highlight that subtle

differences not tapped by more typically-used math equivalence items remain important for

assessing students' knowledge of equivalence. Indeed, with the exception of defining the equal

sign, the majority of past research has focused on items that tap understanding at Levels 1, 2, and

3 of the construct map (e.g., Alibali, 1999; Baroody & Ginsburg, 1983; Li et al., 2008; McNeil &

Alibali, 2005). From an item response theory perspective, the Level 4 items add important

information about learners who have moved beyond the basic levels of equivalence knowledge.

From a practical perspective, this suggests Level 4 items should be included in assessments of

math equivalence knowledge in order to obtain a nuanced picture of student understanding.

The inclusion of Level 4 items also provided empirical evidence for a solid connection

between knowledge of symbolic math equivalence and at least some aspects of formal algebra

(e.g., Alibali et al., 2007; Knuth et al., 2006; MacGregor & Stacey, 1997; Steinberg et al., 1991). Several Level 4 equation-solving items were basic algebra problems with an unknown variable (e.g., $c + c + 4 = 16$). As in prior work (Matthews et al., 2012), these items loaded highly on the math equivalence construct. That is, they fit well with the other items and functioned in predictable ways, even in this cross-section of pre-algebra and algebra students. This provides evidence that developing knowledge of algebra is strongly linked to knowledge of equivalence.

We also generated new findings about the links between equivalence knowledge and interpretation of variables: knowledge of math equivalence was related to students' conceptual interpretations of an algebraic expression that did not explicitly contain the equal sign. For pre-algebra students, the use of mnemonic letters (c and b to stand for the cost of cakes and brownies as opposed to the more traditional x and y) interfered with their ability to conceptually interpret the expression (see McNeil et al., 2010). In contrast, algebra students exhibited a deeper understanding of the expression that was not influenced by the specific letter variables used. Importantly, students' empirically-derived ability estimates on the math equivalence assessment predicted their likelihood of interpreting the algebraic expression correctly, even after controlling for their current math course, the letter variables used, and their age. This lends support to the notion that a nuanced understanding of math equivalence extends to the concept as a whole beyond the use of the formal "=" symbol. It supports the broader idea that students' conception of math equivalence progresses from an operational view to a relational/structural view (e.g., Kieran, 1981; Sfard & Linchevski, 1994) and that where they are in this progression predicts their reasoning about expressions with variables on a formal algebra task.

A final contribution of the current research is to reinforce the benefits of combining quantitative and qualitative methodological approaches in integrative ways. For example, many

of the Level 4 items required a qualitative coding of students' written responses. It was often

insufficient to know whether the student judged the equation as true or false. Rather, we had to

take into account the student's written explanation and to look for evidence of comparative

relational understanding. These coding schemes were influenced heavily by qualitative work in

mathematics education, such as that by Behr et al., (1980) and Carpenter et al. (2003). After the

initial qualitative coding of student responses, a quantitative psychometric approach was applied

(i.e., the Rasch model), which allowed us to obtain empirical estimates of item difficulties and

student abilities. Finally, differences in item difficulties and student ability estimates helped to

identify the items on which students varied in key ways, allowing us to take a closer, qualitative

look at students' errors on those items. Thus, this iterative process not only showed that the

qualitative and quantitative analyses were fully compatible, but also provided greater insight into

the structure of students' knowledge than either approach alone.

Despite the contributions of the current research, there are a number of limitations. First,

although we provided some evidence for the validity of the assessment, we did not include

additional measures that would allow us to assess discriminant validity (e.g., ensuring the

assessment is not measuring a different construct). Further, the lack of additional measures of

algebraic knowledge prevents us from providing a benchmark for assessing the strength of the

correlation of equivalence knowledge with algebraic understanding. In future work, a

comprehensive pretest of algebra knowledge would go much further both in terms of confirming

the differences in skills among the two cohorts and in terms of charting the correlation between

level of algebra proficiency and equivalence knowledge. These issues somewhat limit the

conclusions we can draw, particularly in terms of the assessment's utility in correlational data

analysis. However, to our knowledge, there is currently no other existing psychometrically-

validated criterion measure for assessing knowledge of math equivalence. Our work is intended to push the field on this end, and future work is needed to corroborate our inferences.

Second, based on the previous measurement studies on this assessment (Matthews et al., 2012; Rittle-Johnson et al., 2011), we opted to use a construct-modeling approach with a one-parameter Rasch model. These methodological decisions were justified given our research aims, but we acknowledge that there are additional or alternative techniques that may enhance the measurement development process. For example, it is certainly possible that a two-parameter model would result in better model fit had we gathered a significantly larger sample that would allow us to use such a model. In the future, using one form of the assessment rather than two separate forms would reduce the sample size necessary to use a two-parameter model. Moreover, our analysis also does not allow us to directly compare the item difficulty estimates in this sample to those found in previous elementary-school samples. Thus, although we can examine whether performance in this older group supports the construct map and hypothesized order of difficulties, we cannot make explicit claims regarding the similarity of measurement properties across younger and older samples from different studies. Studies designed explicitly to facilitate vertical scaling across age groups would add more clarity on this end. Such studies could be specifically designed to deal with the fact that items that provide little information for one cohort, because they are very difficult or very easy for that cohort, might provide considerable information for another.

Third, the generalizability of the results remains unknown given several design decisions. We used a convenience sample of pre-algebra and algebra students from two schools within the same geographical region. We did not collect a large array of individual-level demographic characteristics, and this limited our understanding of our sample's representativeness to the

larger population. Future work with more diverse populations who are served with diverse curricula is necessary to get a measure of the generalizability of our finding. Also, we used a cross-sectional design allowing us to note differences between pre-algebra and algebra students. However, longitudinal studies will be necessary to track changes in math equivalence understanding over time. Moreover, we administered the assessment in one shot as part of a measurement research study. Future research should investigate its potential use for formative assessment in real classrooms to identify students with weak understanding and to assess changes in knowledge in response to intervention. Finally, we showed that math equivalence knowledge is related to students' conceptual interpretations of an algebraic expression, but we relied on a single item to assess these interpretations. Future research could examine whether knowledge of math equivalence predicts performance on a more comprehensive assessment of variable understanding.

Given the push to make algebra accessible to all students, it is imperative to measure emerging algebraic knowledge with valid, comprehensive assessments. In the current measurement endeavor, we did just that – we focused on the assessment of math equivalence knowledge beyond elementary school and provided empirical support for the link between knowledge of equivalence and formal algebraic reasoning in middle school students.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education.

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology*, *35*, 127-145. doi:10.1037/0012-1649.35.1.127

Alibali, M. W., Knuth, E. J., Hattikudur, S., McNeil, N. M., & Stephens, A. C. (2007). A longitudinal examination of middle school students' understanding of the equal sign and equivalent equations. *Mathematical Thinking and Learning, 9*, 221–247. doi:10.1080/10986060701360902

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. *Elementary School Journal, 84*, 199-212. doi:10.1086/461356

Behr, M., Erlwanger, S., & Nichols, E. (1980). How children view the equal sign. *Mathematics Teaching, 92*, 13-15.

Blanton, M., Stephens, A., Knuth, E., Gardiner, A. M., Lsler, I., & Kim, J. S. (2015). The development of children's early algebraic thinking: The impact of a comprehensive early algebra intervention in third grade. *Journal for Research in Mathematics Education, 46*, 39-87. doi:10.5951/jresematheduc.46.1.0039

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahway, NJ: Erlbaum.

Booth, J. L., & Davenport, J. D. (2013). The role of problem representation and feature knowledge in algebraic equation-solving. *Journal of Mathematical Behavior, 32*, 415-423. doi:10.1016/j.jmathb.2013.04.003

Carpenter, T. P., Franke, M. L. & Levi, L. (2003). Thinking mathematically: Integrating

    arithmetic and algebra in elementary school. Portsmouth, NH: Heinemann.

Charles, R. I. (2005). Big ideas and understandings as the foundation for elementary and middle

    school mathematics. *National Council for Supervisors of Mathematics: Journal of*

    Mathematics Education Leadership, 8, 9-24.

Chesney, D., McNeil, N. M., Brockmole, J. R., & Kelley, K. (2013). An eye for relations: Eye-

    tracking indicates long-term negative effects of operational thinking on understanding of

    math equivalence. *Memory and Cognition, 41*, 1079-1095. doi:10.3758/s13421-013-

    0315-8

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A

    foundation for algebra. *Teaching Children Mathematics*, *6*, 232-236.

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal

    role of prior knowledge. *Journal of Educational Psychology, 108,* 82-97.

    doi:10.1037/edu0000053

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When

    conceptual instruction prior to problem solving improves mathematical knowledge.

    British Journal of Educational Psychology, 84, 502-519. doi:10.1111/bjep.12035

Fyfe, E. R., & McNeil, N. M, & Borjas, S. (2015). Benefits of "concreteness fading" for

    children's mathematics understanding. *Learning and Instruction, 35,* 104-120.

    doi:10.1016/j.learninstruc.2014.10.004

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during

    exploratory mathematics problem solving: Prior knowledge matters. *Journal of*

    Educational Psychology, 104, 1094-1108. doi:10.1037/a0028389

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Herscovics, N., & Kieran, C. (1980). Constructing meaning for the concept of equation. *Mathematics Teacher, 73*, 572-580.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics, 27*, 59-78.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38,* 258-288.

Jones, I., Inglis, M., Gilmore, C., & Dowens, M. (2012). Substitution and sameness: Two components of a relational conception of the equals sign. *Journal of Experimental Child Psychology*, *113*, 166–176. doi:10.1016/j.jecp.2012.05.003

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12*, 317-326. doi:10.1007/BF00311062

Kinzel, M. T. (1999). Understanding algebraic notation from the students' perspective. *Mathematics Teacher, 92*, 436–442.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 37*, 297-312.

Knuth, E. J., Alibali, M. W., McNeil, N. M., Weinberg, A., & Stephens, A. C. (2005). Middle school students' understanding of core algebraic concepts: Equality and variable. *International Reviews on Mathematical Education, 37*, 1–9.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.). New York, NY: Springer-Verlag.

Küchemann, D. (1978). Children's understanding of numerical variables. *Mathematics in School, 7*, 23–26.

Li, X., Ding, M., Capraro, M. M., & Capraro, R. M. (2008). Sources of differences in children's understandings of mathematical equality: Comparative analysis of teacher guides and student texts in China and the United States. *Cognition and Instruction, 26*, 195-217. doi: 10.1080/07370000801980845

Linacre, J. M. (2013). *Winsteps Version 3.80.1 Computer Software*. Beaverton, Oregon. Winsteps.com. Available from http://www.winsteps.com.

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program user's guide*. Beaverton, Oregon: Winsteps.com. http://www.winsteps.com/winman/index.htm?copyright.htm

Linchevski, L., & Herscovics, N. (1996). Crossing the cognitive gap between arithmetic and algebra: Operating on the unknown in the context of equations. *Educational Studies in Mathematics, 30*, 39-65.

Lindvall, C. M., & Ibarra, C. G. (1980). Incorrect procedures used by primary grade pupils in solving open addition and subtraction sentences. *Journal for Research in Mathematics Education, 11*, 50-62.

Lucariello, J., Tine, M. T., & Ganley, C. M. (2014). A formative assessment of students' algebraic variable misconceptions. *Journal of Mathematical Behavior, 33*, 30-41. doi:10.1016/j.jmathb.2013.09.001

MacGregor, M., & Stacey, K. (1997). Students' understanding of algebraic notation. *Educational Studies in Mathematics, 33*, 1-19.

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104*, 1–21. doi:10.1016/j.jecp.2008.08.004

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43*, 316-350.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*, 1-17. doi:10.1111/j.1467-8624.2005.00884.x

McNeil, N. M., Fyfe, E. R., & Dunwiddie, A. E. (2015). Arithmetic practice can be modified to promote understanding of mathematical equivalence. *Journal of Educational Psychology, 107,* 423-436. doi:10.1037/a0037687

McNeil, N. M., Rittle-Johnson, B., Hattikudur, S., & Petersen, L. A. (2010). Continuity in representation between children and adults: Arithmetic knowledge hinders undergraduates' algebraic problem solving. *Journal of Cognition and Development, 11*, 437–457. doi:10.1080/15248372.2010.516421

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing $4 = 2 + 2$: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development, 82*, 1620-1633. doi:10.1111/j.1467-8624.2011.01622.x

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-school students' understanding of the equal sign: The books

they read can't help. *Cognition and Instruction, 24*, 367–385.

doi:10.1207/s1532690xci2403_3

McNeil, N. M., Weinberg, A., Hattikudur, S., Stephens, A. C., Asquith, P., Knuth, E. J., & Alibali, M. W. (2010). A is for apple: Mnemonic symbols hinder the interpretation of algebraic expressions. *Journal of Educational Psychology*, *102*, 625–634. doi:10.1037/a0019105

Mix, K. (1999). Preschoolers' recognition of numerical equivalence: Sequential sets. *Journal of Experimental Child Psychology, 74*, 309-332. doi:10.1006/jecp.1999.2533

Molina, M., & Ambrose, R. (2008). From an operational to a relational conception of the equal sign: Third graders' developing algebraic thinking. *Focus on Learning Problems in Mathematics*, *30*, 61-80.

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Authors.

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

National Governors Association Center for Best Practices and the Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. Washington, DC: NGA Center & CCSSO. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development, 6*, 449-468. doi:10.1016/0885-2014(91)90049-J

Powell, S. R., & Fuchs, L. S. (2010). Contribution of equal-sign instruction beyond word-problem tutoring for third-grade students with mathematics difficulty. *Journal of Educational Psychology*, *102*, 381–394. doi:10.1037/a0018447

Powell, S. R., Kearns, D. M., & Driver, M. K. (2016). Exploring the connection between arithmetic and prealgebraic reasoning at first and second grade. *Journal of Educational Psychology, 108,* 943-959. doi:10.1037/edu0000112

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press. (Original work published 1960)

Renwick, E. (1932). Children's misconceptions concerning the symbols for mathematical equality. *The British Journal of Educational Psychology*, *2*, 173-183.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*, 1-15. doi:10.1111/j.1467-8624.2006.00852.x

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*, 175-189. doi:10.1037/0022-0663.91.1.175

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

Rosnick, P. (1981). Some misconceptions concerning the concept of variable. *Mathematics Teacher, 74*, 418-420.

Seo, K. H., & Ginsburg, H. P. (2003). "You've got to carefully read the math sentence…": Classroom context and children's interpretations of the equal sign. In A. J. Baroody & A.

Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sfard, A., & Linchevski, L. (1994). The gains and pitfalls of reification: The case of algebra. *Educational Studies in Mathematics, 26*, 191-228.

Sherman, J., & Bisanz, J. (2009). Equivalence in symbolic and nonsymbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology, 101*, 88-100. doi:10.1037/a0013156

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 25-40.

Steinberg, R. M., Sleeman, D. H., & Ktorza, D. (1991). Algebra students' knowledge of equivalence of equations. *Journal for Research in Mathematics Education, 22*, 112–121.

Stephens, A., Blanton, M., Knuth, E., Isler, I., & Gardiner, A. M. (2015). Just say yes to early algebra! *Teaching Children Mathematics, 22*, 92-101. doi:10.5951/teacchilmath.22.2.0092

Van Amerom, B. A. (2003). Focusing on informal strategies when linking arithmetic to early algebra. *Educational Studies in Mathematics, 54*, 63-75.

Weaver, J. F. (1973). The symmetric property of the equality relation and young children's ability to solve open addition and subtraction sentences. *Journal for Research in Mathematics Education, 4*, 45–56. doi:10.2307/749023

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116. doi:10.1111/j.1745-3984.1977.tb00039.x

Wright, B. D., & Linacre, J. M. (1994). *Reasonable mean-square fit values*. Rasch Measurement Transactions, 8, 370. http://www.rasch.org/rmt/rmt83b.htm.

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

Erlbaum.

Appendix: Math Equivalence Assessment Items

#	Level	Form 2 Item	Form 1 Item
		Equation-Structure Items (ST)	
1.[A]	1	8 = 5 + 10 (true or false?)	8 = 5 + 10 (true or false?)
2.	2	3 = 3 (true or false?)	8 = 8 (true or false?)
3.	2	6 = 6 + 0 (true or false?)	4 = 4 + 0 (true or false?)
4.	2	7 = 3 + 4 (explain why true)	8 = 5 + 3 (explain why true)
5.[A]	3	31 + 16 = 16 + 31 (true or false?)	31 + 16 = 16 + 31 (true or false?)
6.	3	7 + 6 = 6 + 6 + 1 (true or false?)	3 + 1 = 1 + 1 + 2 (true or false?)
7.[A]	3	6 + 4 = 5 + 5 (explain why true)	6 + 4 = 5 + 5 (explain why true)
8.	4	67 + 86 = 68 + 85 (explain why true)	89 + 44 = 87 + 46 (explain why true)
9.	4	8 + 2 + __ = 10 + __	1 + 9 + __ = 10 + __
10.	4	If 17 + 12 = 29, does 17 + 12 + 8 = 29 + 8? (explain why)	If 25 + 14 = 39, does 25 + 14 + 7 = 39 + 7? (explain why)
11.	4	If 2 x 3 = 6, does 2 x 3 x 4 = 6 x 4?	If 3 x 5 = 15, does 3 x 5 x 6 = 15 x 6?
12.[A]	4	If 76 + 49 = 121, does 76 + 49 – 9 = 121 – 9? (explain without subtracting)	If 76 + 49 = 121, does 76 + 49 – 9 = 121 – 9? (explain without subtracting)
13.	4	2 x __ = 58, 8 x 2 x __ = 8 x 58 (why same number goes in the blanks)	3 x __ = 45, 7 x 3 x __ = 7 x 45 (why same number goes in the blanks)
		Equal-Sign Items (ES)	
14.	1	6 + 4 (identify an equal pair)	3 + 6 (identify an equal pair)
15.	2	5 cents __ one nickel (select symbol that shows they are the same)	10 cents __ one dime (select symbol that shows they are the same)
16.	3	Is "the same as" a good definition of the equal sign?	Is "two amounts are the same" a good definition of the equal sign?
17.[A]	4	Which is best definition of equal sign?	Which is best definition of equal sign?
18.[A]	4	1 quarter = 25 pennies (define)	1 quarter = 25 pennies (define)
19.[A]	4	What does the equal sign mean?	What does the equal sign mean?
		Open Equation-Solving Items (OE)	
20.	1	4 + __ = 8	__ + 5 = 9
21.	2	8 = 6 + __	7 = __ + 3
22.[A]	3	3 + 6 = 8 + __	3 + 6 = 8 + __
23.	3	__ + 2 = 6 + 4	5 + __ = 6 + 2
24.[A]	3	7 + 6 + 4 = 7 + __	7 + 6 + 4 = 7 + __
25.*	3	8 + __ = 8 + 6 + 4	__ = 8 + 5 + 9
26.	3	6 – 4 + 3 = __ + 3	8 + 5 – 3 = 8 + __
27.[A]	4	67 + 84 = __ + 83	67 + 84 = __ + 83
28.	4	43 + __ = 48 + 76	__ + 55 = 37 + 54
29	3	10 = z + 6	13 = n + 5
30.[A]	4	c + c + 4 = 16	c + c + 4 = 16
31.	4	m + m + m = m + 12	z + z + z = z + 8

Note. [A] indicates anchor items, * indicates Form 1 item was intended to be __ + 9 = 8 + 5 + 9, but misprint renders it Level 2. All "Levels" were assigned *a priori* and hypothesized based on the construct map.

Table 1

Construct Map for Math Equivalence Knowledge

Level	Description	Equation Structure		Sample Items
4 Comparative Relational	Successful with a variety of equation structures by comparing expressions on the two sides of the equal sign. Generate a relational definition of the equal sign.	Operations on both sides with multi-digit numbers or with multiple instances of a variable.	ST	67 + 86 = 68 + 85 (true or false? explain why)
			EQ	What does the equal sign mean?
			OE	__ + 55 = 37 + 54
3 Basic Relational	Successful with equation structures with operations on both sides of the equal sign. Recognize a relational definition of the equal sign as correct.	Operations on both sides: $a + b = c + d$ $a + b - c = d + e$	ST	31 + 16 = 16 + 31 (true or false?)
			EQ	Is "the same as" a good definition of equal sign?
			OE	5 + __ = 6 + 2
2 Flexible Operational	Successful with equation structures that are compatible with an operational view of the equal sign.	Operations on the right side: $c = a + b$ No operations: $a = a$	ST	4 = 4 + 0 (true or false?)
			EQ	10 cents __ one dime (select correct symbol)
			OE	7 = __ + 3
1 Rigid Operational	Successful with an operations-equals-answer equation structure. Generate an operational definition of the equal sign.	Operations on the left side: $a + b = c$	ST	5 + 2 = 7 (true or false?)
			EQ	Identify a pair equal to 3 + 6
			OE	__ + 5 = 9

Note. Table adapted from Rittle-Johnson et al. (2011, p. 87). Success includes the ability to solve, evaluate, and encode equations of a particular structure. ST = Equation-Structure Items, EQ = Equal-Sign Items, OE = Open Equation-Solving Items.

Table 2

Examples of Students' Interpretations of the Algebraic Expression

Interpretations of the expression 4c + 3b
Correct Interpretations
"The total cost of cakes and brownies"
"The amount of money you'll pay for the cakes and brownies"
"The price of four cakes plus the price of three brownies"
"Four times the cost of one cake plus three times the cost of one brownie"
Incorrect Interpretations
"Four cakes plus three brownies"
"The number of cakes and brownies I bought"
"It means you do 4 times c and 3 times b"
"That is the equation to find the answer"

Table 3

Item Statistics for Mathematical Equivalence Assessment

Construct Component	#	Hypothesized Level	Item Name	Item Example	Mean Accuracy	Observed item location (δ, in Logits)	Standard error of δ estimate	Infit MSQ	Outfit MSQ
Structure	1[A]	1	L1_Structure1_V1V2	8 = 5 + 10 (True or False)	0.96	-1.57	0.36	0.9544	0.6267
Structure	2	2	L2_Structure2_V1	8 = 8 (True or False)	0.97	-2.07	0.61	1.0946	0.5365
Structure	3	2	L2_Structure3_V1	4 = 4 + 0 (True or False)	0.96	-1.74	0.54	1.0287	1.3742
Structure	4	2	L2_Structure4_V1	8 = 5 + 3 (Explain why True)	0.97	-2.03	0.62	0.7506	0.1427[B]
Structure	5[A]	3	L3_Structure5_V1V2	31 + 16 = 16 + 31 (True or False)	0.96	-1.57	0.36	0.9006	0.5156
Structure	6	3	L3_Structure6_V1	3 + 1 = 1 + 1 + 2 (True or False)	0.95	-1.25	0.45	1.0937	1.5785
Structure	7[A]	3	L3_Structure7_V1V2	6 + 4 = 5 + 5 (Explain why True)	0.9	-0.37	0.25	0.7184	0.542
Structure	8	4	L4_Structure8_V1	89 + 44 = 87 + 46 (Explain why True)	0.59	2.02	0.22	1.0444	1.153
Structure	9	4	L4_Structure9_V1	1 + 9 + □ = 10 + □ (Explain which numbers could go in box)	0.51	2.45	0.22	0.9688	0.8884
Structure	10	4	L4_Structure10_V1	If 25 + 14 = 39, does 25 + 14 + 7 = 39 + 7? (explain why)	0.59	2.02	0.22	0.8435	0.7927
Structure	11	4	L4_Structure11_V1	If 3 x 5 = 15, does 3 x 5 x 6 = 15 x 6? (explain why)	0.49	2.55	0.22	0.9907	0.9444
Structure	12[A]	4	L4_Structure12_V1V2	If 76 + 49 = 121, does 76 + 49 – 9 = 121 – 9? (explain why without subtracting)	0.47	2.74	0.16	0.8756	0.8002
Structure	13	4	L4_Structure13_V1	3 x __ = 45, 7 x 3 x __ = 7 x 45 (explain why the same number goes in the blanks)	0.15	4.89	0.31	0.9575	0.6901
Equal-Sign	14	1	L1_EqualSign14_V1	3 + 6 (Identify an equal pair)	0.92	-0.73	0.38	1.2629	2.6169
Equal-Sign	15	2	L2_EqualSign15_V1	10 cents __ one dime (select symbol that shows they are the same)	0.96	-1.74	0.54	1.0759	0.4256[B]
Equal-Sign	16	3	L3_EqualSign16_V1	Is "two amounts are the same" a good definition of the equal sign?	0.93	-0.89	0.4	0.9976	0.502
Equal-Sign	17	4	L4_EqualSign17_V1	Which (of three options) is the best definition of the equal sign?	0.67	1.56	0.23	1.0724	1.1386

Construct Component	#	Hypothesized Level	Item Name	Item Example	Mean Accuracy	Observed item location (δ, in Logits)	Standard error of δ estimate	Infit MSQ	Outfit MSQ
Equal-Sign	18[A]	4	L4_EqualSign18_V1V2	1 quarter = 25 pennies (define equal sign in this context)	0.73	1.2	0.17	1.1439	1.0716
Equal-Sign	19[A]	4	L4_EqualSign19_V1V2	What does the equal sign mean?	0.62	1.88	0.16	1.0508	1.0247
Solving	20	1	L1_Solve20_V1	__ + 5 = 9	0.98	-2.52	0.74	1.1714	7.6052[B]
Solving	21	2	L2_Solve21_V1	7 = __ + 3	0.99	-3.26	1.02	0.9759	0.1575[B]
Solving	22[A]	3	L3_Solve22_V1V2	3 + 6 = 8 + __	0.97	-2.04	0.43	0.8617	1.5427[B]
Solving	23	3	L3_Solve23_V1	5 + __ = 6 + 2	0.96	-1.74	0.54	0.8749	0.2063[B]
Solving	24[A]	3	L3_Solve24_V1V2	7 + 6 + 4 = 7 + __	0.9	-0.31	0.24	1.0839	1.9299[B]
Solving	25	2	L2_Solve25_V1	__ = 8 + 5 + 9	0.94	-1.06	0.42	1.1763	5.1268[B]
Solving	26	3	L3_Solve26_V1	8 + 5 – 3 = 8 + __	0.9	-0.46	0.35	0.9169	0.9855
Solving	27[A]	4	L4_Solve27_V1V2	67 + 84 = __ + 83	0.81	0.61	0.19	0.8555	0.6237
Solving	28	4	L4_Solve28_V1	__ + 55 = 37 + 54	0.75	0.98	0.25	1.0703	1.9032[B]
Solving	29	3	L3_Solve29_V1	13 = n + 5	0.92	-0.73	0.38	1.1461	1.0362
Solving	30[A]	4	L4_Solve30_V1V2	c + c + 4 = 16	0.69	1.49	0.17	1.018	1.1409
Solving	31	4	L4_Solve31_V1	z + z + z = z + 8	0.97	1.77	0.23	0.9584	0.8911

Note. The data table is based on collapsing the data from the two forms of the revised assessment, with example items from one of the forms. [A] indicates anchor items. [B] Indicates items with outfit scores out of the desired range. These items tended to be items for which the current sample demonstrated ceiling effects.

Table 4

Confirmatory Factor Analysis Exploring Unidimensionality

	Version 1		Version 2	
Measure	1-Factor Model	3-Factor Model	1-Factor Model	3-Factor Model
RMSEA	.067	.062	.045	.043
RMSEA CI	(.056, .077)	(.051, .073)	(.038, .053)	(.035, .051)
CFI	.725	.762	.762	.735
Chi-Sq	652.831	619.710	637.467	613.687
df	df = 434	df = 432	df=434	df = 321

Table 5

Probabilities of Success Based on Item Difficulty Estimates and Student Ability Estimates

Item	Hypothesized Difficulty Level	Difficulty Estimate, δ	Probability of Success	
			Pre-Algebra Mean Ability Estimate, $\theta = 2.3$	Algebra Mean Ability Estimate, $\theta = 2.9$
$7 = \_\_ + 3$	2	-3.26	0.996	0.998
$\_\_ + 5 = 9$	1	-2.52	0.992	0.996
$4 = 4 + 0$ (true or false)	2	-1.74	0.983	0.990
$13 = n + 5$	3	-0.73	0.954	0.974
If $25 + 14 = 39$, does $25 + 14 + 7 = 39 + 7$? (explain why)	4	2.02	0.574	0.711
$3 \times \_\_ = 45$, $7 \times 3 \times \_\_ = 7 \times 45$ (explain why the same number goes in the blanks)	4	4.89	0.069	0.119

Note. Entries in the two rightmost columns represent the probabilities that the average student of a given ability estimate (2.3 or 2.9) will answer an item of a given difficulty estimate correctly. Difficulty level is the hypothesized level based on the construct map. Difficulty estimate is the empirically-derived difficulty estimates based on the Rasch analysis.
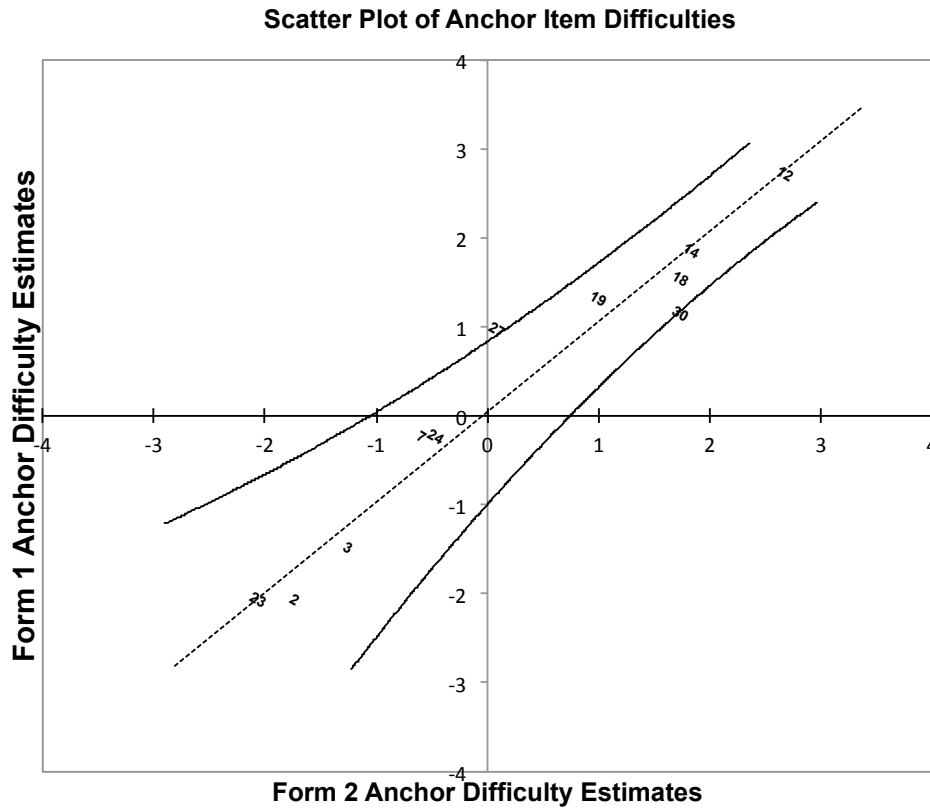
Scatter Plot of Anchor Item Difficulties

Figure 1. Scatterplot of anchor item difficulties for Form 1 vs. Form 2. Plots anchor item difficulty estimates in logits for both Forms of the assessment. Best fit line has a slope near one and an intercept near zero, indicating that the scales for each form are interchangeable.
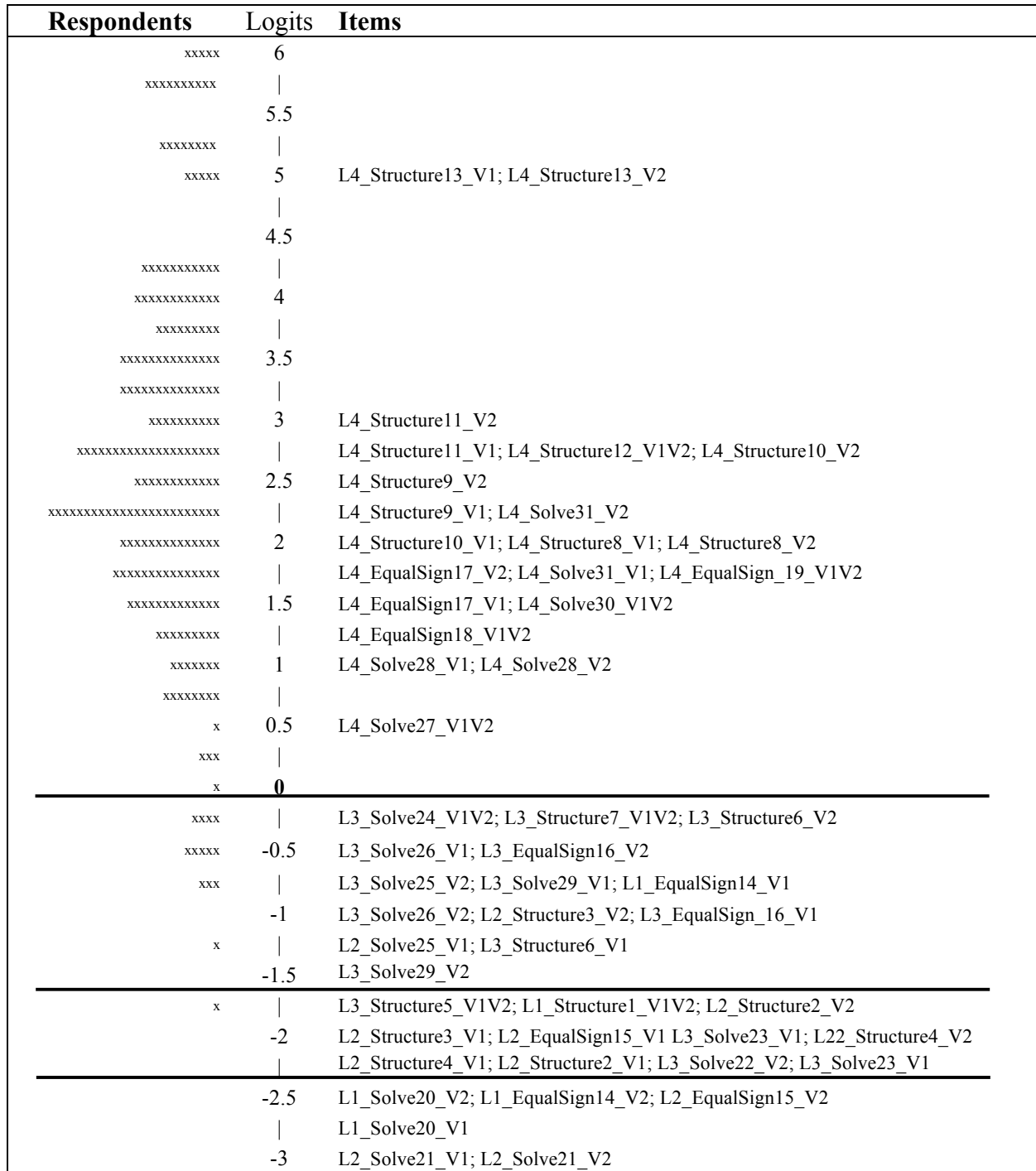
Respondents	Logits	Items
xxxxx	6	
xxxxxxxxxx		
	5.5	
xxxxxxxx		
xxxxx	5	L4_Structure13_V1; L4_Structure13_V2
	4.5	
xxxxxxxxxxx		
xxxxxxxxxxxx	4	
xxxxxxxxx		
xxxxxxxxxxxxxx	3.5	
xxxxxxxxxxxxxx		
xxxxxxxxxx	3	L4_Structure11_V2
xxxxxxxxxxxxxxxxxxxxx		L4_Structure11_V1; L4_Structure12_V1V2; L4_Structure10_V2
xxxxxxxxxxxx	2.5	L4_Structure9_V2
xxxxxxxxxxxxxxxxxxxxxxxx		L4_Structure9_V1; L4_Solve31_V2
xxxxxxxxxxxxx	2	L4_Structure10_V1; L4_Structure8_V1; L4_Structure8_V2
xxxxxxxxxxxxxx		L4_EqualSign17_V2; L4_Solve31_V1; L4_EqualSign_19_V1V2
xxxxxxxxxxxx	1.5	L4_EqualSign17_V1; L4_Solve30_V1V2
xxxxxxxx		L4_EqualSign18_V1V2
xxxxxx	1	L4_Solve28_V1; L4_Solve28_V2
xxxxxxxx		
x	0.5	L4_Solve27_V1V2
xxx		
x	**0**	
xxxx		L3_Solve24_V1V2; L3_Structure7_V1V2; L3_Structure6_V2
xxxxx	-0.5	L3_Solve26_V1; L3_EqualSign16_V2
xxx		L3_Solve25_V2; L3_Solve29_V1; L1_EqualSign14_V1
	-1	L3_Solve26_V2; L2_Structure3_V2; L3_EqualSign_16_V1
x		L2_Solve25_V1; L3_Structure6_V1
	-1.5	L3_Solve29_V2
x		L3_Structure5_V1V2; L1_Structure1_V1V2; L2_Structure2_V2
	-2	L2_Structure3_V1; L2_EqualSign15_V1 L3_Solve23_V1; L22_Structure4_V2
		L2_Structure4_V1; L2_Structure2_V1; L3_Solve22_V2; L3_Solve23_V1
	-2.5	L1_Solve20_V2; L1_EqualSign14_V2; L2_EqualSign15_V2
		L1_Solve20_V1
	-3	L2_Solve21_V1; L2_Solve21_V2

Figure 2. Wright map for the math equivalence assessment. Each x on the left represents one student. Each entry on the right represents one item. Item entries name the *hypothesized* difficulty level (e.g., L4), the item type (e.g., Structure), the item number (e.g., 13), and the form number (e.g., V1). The numbers on the vertical axis represent item difficulty and student ability estimates in logits. The horizontal lines are for visual, descriptive purposes only – the construct is theorized to be continuous.