Understanding the Equals Sign as a Gateway to Algebraic Thinking

Percival G. Matthews, University of Notre Dame; Bethany Rittle-Johnson, Vanderbilt University; University; and Roger S. Taylor, State University of New York at Oswego;

Acknowledgements

This research was supported with funding from the National Science Foundation (NSF) grant DRL-0746565. It was also supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B040110 to Vanderbilt University. The opinions expressed are those of the authors and do not represent the views of the NSF or of the U.S. Department of Education. A special thanks to the students and teachers who participated in this research. Thanks to Caroline Cochrane-Braswell, Holly Harris, Robin Hill, Caitlyn Majeika, Laura McClean, Jadie Pinkston, Ann Simonson, Christine Tanner, Kristen Tremblay, and Ross Troseth for help in reviewing the research literature and collecting and coding the data. Thanks to E. Warren Lambert of the Vanderbilt Kennedy Center and Ryan Kettler of the Learning Sciences Institute at Vanderbilt for statistical help. A special thank you to Maria Blanton, Tom Carpenter, Analucia Schliemann and Jon Star for their expert opinions on the items and construct map.

Correspondence concerning this article should be addressed to Bethany Rittle-Johnson, 230 Appleton Place, Peabody #0552, Nashville, TN 37203, 615-322-8301 (voice), b.rittle-johnson@vanderbilt.edu

Abstract

Knowledge of mathematical equivalence, the principle that two sides of an equation are interchangeable, is a foundational concept of mathematics that serves as a key link between arithmetic and algebra. This knowledge develops throughout elementary and middle school. Unfortunately, measurement issues have limited our abilities to chart the variability in children's developing conceptions of mathematical equivalence. In particular, the diversity of items used by different researchers has made it difficult to compare results across studies. In this study, we used a construct modeling approach to unify these diverse measures into a single instrument designed to measure equivalence knowledge. Our new instrument contributes to the field by a) placing commonly used items on the same metric, making them truly commensurable; b) expanding our abilities to model the variability of student's knowledge of equivalence; and c) showing explicitly that a subset of typical algebraic problems loads strongly on the equivalence knowledge construct.

Understanding the Equals Sign as a Gateway to Algebraic Thinking

Mathematical equivalence is a foundational concept of algebraic thinking that serves as a key link between arithmetic and algebra (Baroody & Ginsburg, 1983; Carpenter, Franke, & Levi, 2003; Kieran, 1981; Knuth, Stephens, McNeil, & Alibali, 2006; MacGregor & Stacey, 1997). Typically represented by the '=' symbol, mathematical equivalence, also called mathematical equality, is the principle that two sides of an equation are interchangeable. Understanding of mathematical equivalence requires *relational* thinking: realizing that the equals sign symbolizes the sameness of the expressions or quantities on each side of an equation (Baroody & Ginsburg, 1983; Behr, 1980; Carpenter et al., 2003; McNeil & Alibali, 2005a). There is a general consensus that knowledge of this concept supports greater algebraic competence, including equation-solving skills and algebraic reasoning (Jacobs, Franke, Carpenter, Levi, & Battey, 2007; Kieran, 1992; Knuth et al., 2006; National Research Council, 1998; Steinberg, Sleeman, & Ktorza, 1991). Moreover, because algebra is an important gateway not only into higher mathematics, but also into higher education more generally, the importance of building high-quality understanding of mathematical equivalence is of critical importance (Adelman, 2006; Moses & C. E. Cobb, 2001).

Unfortunately, numerous studies point to the difficulties that American elementary and middle school children have understanding equivalence (e.g., Alibali, 1999; Behr, 1980; Cobb, 1987; Falkner et al., 1999; Jacobs et al., 2007; Li et al., 2008; McNeil, 2007; Perry, 1991; Powell & Fuchs, 2010; Rittle-Johnson, 2006; Rittle-Johnson & Alibali, 1999; Weaver, 1973). Although elementary school students have some basic understanding of what it means for quantities to be equal, this understanding is often not linked to the equals sign (Baroody, Lai & Mix, 2005; Perry, Church & Goldin-Meadow, 1988; Sherman & Bisanz, 2009). Instead of viewing the equals sign as a symbol expressing the interchangeability of two sides of an equation, children often interpret the equals sign as an operator signal that means ‘‘adds up to’’ or ‘‘gets the answer’’ (e.g., Baroody & Ginsburg, 1983; McNeil & Alibali 2005b). This *operational* view of the equals sign can impede development of a *relational* view of the equals sign necessary for deeper understanding of mathematical equivalence (Kieran, 1981; McNeil & Alibali, 2005a).

In explaining the sources of these difficulties, many accounts blame current educational practices. These accounts cite children’s frequent exposure to problems in standard operations-equals-answer format (e.g., 4 + 5 = 9) for the development of an operational view of equivalence. Notably, this format is compatible with an operational view of equivalence, and it is thought to be the most frequent context in which American elementary school children see the equals sign (McNeil, Grandau, Knuth Alibali, Stephens, Hattikudur & Krill, 2006; Li, Ding, Capraro & Capraro, 2008). Because the standard format is seen repeatedly, student often develop schemas which interpret the equals sign operationally, even when such interpretations are not appropriate (e.g., Falkner et. al, 1999; McNeil & Alibali, 2005b; McNeil, 2008) This operational view of the equals sign often persists for many years, serving as an obstacle to the development of flexible problem solving skills and algebraic reasoning (Knuth et al., 2006; McNeil et al., 2006; Steinberg et al., 1991). For example, most students in Grades 1 to 6 solve problems like 8 + 4 = □ + 5 incorrectly, writing 12 or 17 in the box (Falkner, Levi & Carpenter, 1999). Similarly, most American elementary-school children reject closed equations not in a standard “a + b = c” format (e.g., they consider equations such as 3 = 3 and 7 + 6 = 6 + 6 + 1 as false or nonsensical) (Baroody & Ginsburg, 1983; Behr et. al., 1980; Falkner et al, 1999).

Despite the progress that has been made in detailing the sources and signs of children's misconceptions of equivalence, two important measurement issues remain significant obstacles to advancing our understanding of how children's knowledge of equivalence progresses. First, researchers have not used measures with proven reliability and validity when investigating mathematical equivalence (Hill & Shih, 2009; Authors, in press). Second, the measures of equivalence knowledge that are currently in use have tended to be incommensurable. That is to say that a) different researchers have used different classes of items to measure what is taken to be the same construct; and b) this diversity of independently employed measurement items has led to a situation in which it is difficult to compare results obtained across studies. Moreover, to date it has not been possible to order the range of diverse assessment items in a hierarchy of difficulty that might eventually help chart a typical developmental progression.

The Current Study

The current study is part of a project aimed at tackling these issues by unifying diverse items into a single assessment designed to measure a cohesive construct of mathematical equivalence knowledge. We have attempted to address the issues of validity and reliability in another paper that explains the technical details of our measurement development process (Authors, in press). The present paper uses the same methodology as our earlier paper, but instead focuses on how an assessment that includes each of several diverse measurement items can add resolution to our picture of children's understandings of mathematical equivalence. It also employs a new data set, so simultaneously serves as a partial replication and extension of our earlier, psychometrically oriented study.

Measures Currently in Use

In order to highlight the strengths of our approach, we begin by first providing a brief overview of the measures currently used in the field. A review of the literature showed that research on the topic has primarily employed three different classes of equivalence tasks: (1) *open equation-solving items*, such as 8 + 4 = □ + 5 (e.g., Alibali, 1999; Behr, 1980; Jacobs et al., 2007; McNeil, 2007; Matthews & Rittle-Johnson, 2009; Perry, 1991; Powell & Fuchs, 2010; Rittle-Johnson, 2006; Weaver, 1973), (2) *equation-structure items*, such as deciding if 3 + 5 = 5 + 3 is true or false (e.g., Baroody & Ginsburg, 1983; Behr, 1980; Falkner et al., 1999; Molina & Ambrose, 2006; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003), and (3) *equal-sign-definition* items, such as asking children to provide an explicit verbal definition of what the equals sign mean (Behr,1980; Ginsburg, 1977; Knuth et. al., 2006; McNeil et. al. 2006, Seo & Ginsburg, 2003). Though somewhat different in form, each of these classes of items has been accepted as tapping children's knowledge of mathematical equivalence, providing a prima facie justification for our hypothesis that they might be unified and mapped onto the same measurement scale.

Currently, many studies have privileged one or another of these classes of items over the others as indices of equivalence knowledge. For instance, some have used open equations as the sole experimental measure of students' understanding of equivalence (e.g., McNeil, 2007; Weaver, 1973); others, such as Behr (1980), have focused primarily on equation-structure items; and still others (e.g., Knuth et. al., 2006; McNeil & Alibali 2005a) have focused primarily on whether or not students verbally provided a relational interpretation of the equals sign as the primary indicator of sophistication with mathematical equivalence. Researchers who have employed multiple classes of items have tended to compartmentalize them, analyzing different classes as parts of separate subscales (e.g., Jacobs et al, 2007; McNeil & Alibali, 2005b; Powell & Fuchs, 2010; Rittle-Johnson, 2006; Seo & Ginsburg, 2003). Unfortunately, to date there has not been possible to directly compare across these subscales.

This lack of commensurability for different classes of items across studies is unfortunate. We are left unaware of the relative difficulties of commonly used items or the typical order in which competence is gained. This is especially noteworthy given that several authors have suggested that context can play an important role in the degree to which students' knowledge of equivalence is elicited (Cobb 1987; McNeil & Alibali 2005a; McNeil & Alibali 2005b; Seo & Ginsburg 2003; Weaver 1973). Each of the different classes of items can be considered as measuring equivalence knowledge in a somewhat different context. Measuring knowledge in one context to the exclusion of others provides an incomplete picture of children's knowledge – hence the need for instruments that employ multiple contexts.

Moreover, most studies have not systematically varied factors that are suspected to influence item difficulty *within* a given class. In particular, the format of an equation seems very influential. As noted above, several studies suggest that items with standard equations (all operations on the left) are easier than those with all operations on the right side of the equals sign (Weaver, 1973). It has also been suggested that equations with all operations on the right side may be less difficult than equations with operations on both sides of the equals sign (Falkner et. al, 1999). We have systematically composed our assessment of items of various formats and presumed difficulties. Below, we detail specific ways in which equation structure is hypothesized to affect item difficulty.

Detailing the Construct

In our efforts to place the different classes of items and different equation structures on the same measurement scale, we utilized Mark Wilson's Construct Modeling approach to measurement development (Wilson, 2003, 2005). This approach begins by positing a *construct map*, which is a representation of the continuum of knowledge levels for the construct under consideration.

The construct map we developed in our earlier work is presented in Table 1, with less sophisticated knowledge represented at the bottom and more advanced knowledge represented at the top (Authors, in press). Note that representative items from each of the different classes of typical items are present at each level of our construct map. The distinctions between the knowledge levels differ primarily in the types of equation formats with which children are expected to be successful.

According to our construct map, students at Level 1 are expected to have success with equations in the standard operations-equals-answer format, but to fail with equations in other formats. At Level 2, students maintain an operational view of the equals sign, but become somewhat more flexible with respect to the types of equation formats that they correctly solve and accept as valid. Students at this level specifically become comfortable with equations with operations only on the right and equations with no operations. At Level 3, students begin to hold a basic relational view, although it coexists with an operational one. Their nascent relational understanding is primarily manifested in their becoming successful with equations that feature operations on both sides (e.g., $4 + 5 + 8 = \square + 8$), and they recognize a relational definition of the equals sign as a good definition. Finally, Level 4 is a comparative relational understanding of equivalence. Students at this level consistently offer relational interpretation of the equals sign.

Moreover, their reasoning need not be tied to specific computations. For example, students with a comparative understanding know that performing the same actions on each side of an equation maintains its equivalence, without needing to engage in full computation.

Although the construct map is presented as having four levels for purposes of conceptual clarity, our model of the construct is continuous. The continuous nature of the model means that the levels should not be interpreted as discrete stages. Knowledge change is expected to follow a gradual and dynamic progression, with less sophisticated knowledge sometimes coexisting and competing with more advanced knowledge (Siegler, 1996). For example, an operational view of equivalence can even be elicited from adults in certain circumstances (McNeil & Alibali, 2005a, 2005b).

Advantages of Combining the Measurement Classes

This paper is concerned with addressing three ways in which combining ostensibly disparate classes of items on a single assessment can contribute to the depth of our understandings about children's knowledge of mathematical equivalence. First, if all of the items are indeed measuring the same construct, then they might all be measured on the same scale. This would result in a detailed and meaningful hierarchy of difficulty for items that are typically analyzed separately.

Second, our methodology has the potential to add significant resolution to our abilities to illustrate the wide range of variability that exists among children's equivalence knowledge. Currently, we are usually limited to largely binary divisions that result from characterizing students according to whether they succeed or fail at a particular class of items. The either/or distinctions between those who succeed or fail to provide relational definitions of the equals sign (or alternatively, to solve open equations or to accept noncanonical equations as valid) may have significant predictive power, but these distinctions cannot be used to resolve the variability *within* the binary groups that result. By aggregating a host of different nonstandard structures and

Table 1

Construct Map for Mathematical Equivalence Knowledge

Level	Description	Core Equation Structure(s)
Level 4: Comparative Relational	Successfully solve and evaluate equations by comparing the expressions on the two sides of the equals sign, including using compensatory strategies and recognizing that performing the same operations on both sides maintains equivalence. Consistently generate a relational definition of equals sign.	Operations on both sides with multi-digit numbers or multiple instances of an unknown: m + m + m = m + 12
Level 3: Basic Relational	Successfully solve, evaluate and encode equation structures with operations on both sides of the equals sign. Recognize a relational definition of the equals sign.	Operations on both sides: a + b = c + d a + b - c = d + e
Level 2: Flexible Operational	Successfully solve, evaluate and encode atypical equation structures that remain compatible with an operational view of the equals sign.	Operations on right: c = a + b No operations: a = a
Level 1: Rigid Operational	Only successful with equations with an operations-equals-answer structure, including solving, evaluating and encoding equations with this structure. Define the equals sign operationally.	Operations on left: a + b = c (including when blank is before the equals sign)

item classes on a continuous hierarchy of difficulty, we stand to map much of the variability in children's knowledge that would otherwise go undetected.

Third, establishing a single scale can help inform our notions of the applicability and reach of mathematical equivalence as underlying different kinds of mathematical reasoning. Such a scale could serve as an anchor for developing new items that also load heavily on the construct of mathematical equivalence. For instance, we hypothesized that using a single scale would allow us to place two classes of items typically considered to require algebraic reasoning on the same scale as other typically used equivalence items: 1) equations involving letter variables – or *literal variables* (e.g., $n + n + n + 2 = 17$, Jacobs et. al., 2007); and 2) those requiring children to reason about how performing the same operation on each side of an equation preserves equivalence (e.g., if we know that $76 + 45 = 121$, can we tell without adding whether or not $76 + 45 – 9 = 121 – 9$?, inspired by Alibali et. al., 2007; Carpenter et. al., 2003; Steinberg et. al., 1990).

It is often argued that early understanding of mathematical equivalence is key for later success in algebra and other higher math, but the evidence supporting such claims is somewhat indirect. Currently, some studies do show that higher knowledge of mathematical equivalence is predictive of children's abilities to solve typical algebraic equations (De Corte & Verschaffel, 1981; Knuth et al., 2006) or to reason about equivalence of equations (Alibali, Knuth, Hattikudur, McNeil, & Stephens, 2007). None of these studies, however, has explicitly attempted to put more advanced algebraic reasoning items on the same measurement scale as typical equivalence items. Establishing typical algebraic items as measures of the construct of mathematical equivalence knowledge would help solidify the link between mathematical equivalence and algebraic reasoning.

METHOD

Participants

Data were collected from 13 second- through sixth-grade classrooms in two suburban, public schools in Tennessee near the end of the school year. Of the students who completed the assessment, 53 were in second grade (23 girls), 46 were in third grade (25 girls), 29 were in fourth grade (14 girls), 59 were in fifth grade (26 girls), and 37 were in sixth grade (16 girls). The mean age was 10.3 years (SD = 1.6; Min = 7.7; Max. = 14.1). The students were largely Caucasian; approximately 2% of students were from minority groups. The schools served a working- to middle-class population, with approximately 23% of students receiving free or reduced lunch.

The schools used the Tennessee Comprehensive Assessment Program (TCAP) as a standardized measure of educational progress (http://www.state.tn.us/education/assessment/achievement.shtml). Students' scores in math and reading on the 2009 TCAP were obtained from student records for 3rd through 6th grade students (the TCAP is not administered to 2nd graders). Most students (71%) had scored at the Advanced level; 24% of students were considered Proficient and 5% were considered Below Proficient.

Test Development Procedure

Past research has used the three primary classes of items described above for measuring mathematical equivalence. Most of the items on our assessment were taken directly from previously published work or created based on items present in those (Baroody & Ginsburg, 1983; Behr, 1980; Carpenter et al., 2003; Jacobs et al., 2007; Knuth et al., 2006; Matthews & Rittle-Johnson, 2009; Rittle-Johnson, 2006; Seo & Ginsburg, 1983; Warren, 2003; Weaver, 1973). Items were classified as Level 1 (rigid operational), 2 (flexible operational), 3 (basic relational) or 4 (comparative relational) based on their equation structures, as outlined in Table 1. On some items, students were asked to explain their reasoning.

We made minor revisions to the original assessment of (Authors, in press) based on empirical evidence of item performance and feedback from a panel of experts in mathematics education research. Five items from each form of the assessment used in Authors (in press) were cut due to weak psychometric properties. Eight items were added to each form, based on the advice of our panel of math education experts. Two of the added items added were simpler ones, asking if 4 = 4 + 0 was true or false and what the equals sign meant in the context of "1 quarter = 25 pennies."

Four additional items were added that tested students' knowledge of the properties of equivalence, which hold that an equivalence relationship remains true as long as an identical operation is performed on both sides of the equals sign (see Figure 1). These types of problems have been cited as addressing the types of thought that underlie formal transformational algebra (Kilpatrick, Swafford, & Findell, 2001; Steinberg et al., 1991). Although these items can be solved by computation, we coded performance based upon children's explanations of how the problem can be answered *without* computation. Answers coded as correct needed to call upon explicit knowledge of the properties of equivalence (e.g., "minus 7 is on both sides so you don't need it").

We also added two items that featured literal variables that are commonly seen in formal algebra. For instance, one asked students to, "Find the value of n," for the equation $n + n + n + 2$

7. *Without subtracting* the 9, can you tell if the statement below is true or false?

76 + 45 = 121 is true.

Is 76 + 45 – 9 = 121 – 9 true or false?

True False Can't tell without subtracting

How do you know?

Figure 1. Sample item probing student knowledge of the properties of equivalence

= 17 (from Jacobs et. al., 2007). These items are important because the use of literal variables — particularly multiple instances of the variable — tests whether students comprehend that a variable represents a specific and constant number value. There was only a single item of this type on the original assessments, but there were three on each of the new forms.

There were two comparable forms of the assessment created using a step-by-step item matching procedure to ensure similarity of content and difficulty across forms (see Authors, in press for details on item matching). In sum, there were 31 items on each form of the assessment – thirteen Level 4 items, ten Level 3 items, five Level 2 items, and three Level 1 items.

Test Administration

The assessment was administered on a whole-class basis by members of the project team. We used a spiraling technique to distribute the two forms of the assessment in each classroom, alternating between handing out the first and second form of the assessment. Completion of the assessment required approximately 45 minutes. Test directions were read aloud for each type of item in 2nd grade classrooms to minimize the possibility that reading level would affect performance. Otherwise, test administration was identical across grade levels.

Scoring

Each item was scored dichotomously (i.e., 0 for incorrect or 1 for correct). For computation items, students received a point for answers within one of the correct answer to allow for minor calculation errors. For the nine explanation items, students received a point if they mentioned the equivalence relation between values on the two sides of the equation (see Table 2). An independent rater coded responses for 20% of the sample, with a mean exact agreement of 0.99 for Form 1 (range .96 to 1.00) and .97 for Form 2 (range .87 to 1.00).

The Rasch Measurement Model

We used a Rasch model along with methods from Classical Test Theory to evaluate the performance of the assessment. Rasch modeling is a one-parameter member of the item response theory (IRT) family (Bond & Fox, 2007). The Rasch model estimates both respondent ability and item difficulty simultaneously, yielding the probability that a particular respondent will answer a particular item correctly (Rasch, 1993; Wright, 1977). We used Winsteps software version 3.68.0.2 to perform all IRT estimation procedures (www.winsteps.com). In addition to providing item and respondent parameters, the Rasch model estimation procedure provides information on the goodness of fit between empirical parameter estimates and the measurement model via infit and outfit values (see Linacre, 2010); infit and outfit values for an item between 0.5 and 1.5 indicate that the item fits well with the other items on the test.

One intuitive description of a Rasch model is as a probabilistic Guttman scalogram. It integrates the difficulty hierarchy of the Guttman model with a bit more flexibility: The Rasch model is a probabilistic one, which is consistent with a model of human understanding that allows for different types of understandings to coexist at the same time (e.g. Siegler, 1996). One of the advantages of the model is that it uses empirical results to place items on a true continuum.

Table 2. Coding Scheme for Select Explanation Items

Item	Sample Correct Responses	Sample Incorrect Answers
What does the equals sign (=) mean?	*“It means the same as”* *“Is equal to or has the same amount”*	*“The answer to the question”* *“The sum”*
Without subtracting the 7, can you tell if the statement is true or false? 56 + 85 = 141 is true. Is 56 + 85 - 7 = 141 – 7 true or false?	*“Because it's subtracting 7 from both sides and 56+85=141 then subtract seven and it'll be equal”* *“Because before you minus seven you have 141 for both”*	*“I did the math”* *“in my head I subtracted and got the same answer”* *“because I looked at both number sentences and they didn't match”*
Without adding 89 + 44, can you tell if the number sentence is true or false? “89 + 44 = 87 + 46”	*“Because all you do is take two from 89 and put two on 44”* *“Because 89-87=2 and they add 2 to 44 so it is even”*	*“Because 89+44=183 and 87+46=133 too”* *“You just have to add the numbers up”*

Hence, our construct map is really only the conceptual skeleton upon which our model is built. Once the empirical data is used to add substance to the skeleton, those data can be used to seamlessly cover all four levels. These empirical estimates, along with the fit scores discussed above, can be used to address our primary concerns about the variability of children’s understandings and about whether or not typically algebraic items can properly be grouped with other items that measure understanding of mathematical equivalence.

Because equivalent groups took the two forms of the assessment, we were able to use a single IRT model to estimate item difficulty and respondent ability across the forms (Kolen &

Brennan, 2004). Several indicators confirmed the equivalence of students who completed the two forms. The distribution of forms was even within each grade level, and groups were also equivalent in mean age (Form1 = 10.3 years, Form 2 = 10.2 years), mean grade (Form 1 = 4.0, Form 2 = 3.9), and on average TCAP math scores (Form 1 = 525, Form 2 = 517).

RESULTS

We first briefly discuss evidence for the reliability and validity of the assessment, which supports that all of the items can be measured on the same scale. Next, we turn to how the instrument specifies the degree of variability among students' levels of equivalence knowledge. Finally, we discuss how the new measure can explicitly show the relation between equivalence knowledge and more advanced algebraic competence.

Evidence for Reliability & Validity – A Single Construct Model

Internal consistency, as assessed by Cronbach's α, was high for both forms of the assessment (Form 1 = .93; Form 2 = .94), providing support for the reliability of the assessment. Multiple measures provided evidence for the validity of our assessment. First, four mathematics education experts who each had over 10 years experience conducting research on elementary-school children's knowledge of algebra rated the items, providing evidence for face validity of the test content. The four experts rated most of the test items as ranging from *important* (rating = 3 of 5) to *essential* (rating = 5 of 5) items for tapping knowledge of equivalence, with a mean rating of 4.1.

Second, we evaluated whether our construct was reasonably characterized as tapping a single dimension. Within an IRT framework, the unidimensionality of a measure is often checked by using a principle components analysis of the residuals (PCA) after fitting the data to the Rasch model (Linacre, 2010). This analysis attempts to partition unexplained variance into coherent factors that may indicate other dimensions. The Rasch model accounted for 59.9% of the variance in our data set. A PCA on the residuals indicated that the largest secondary factor accounted for 2.1% of the total variance (eigenvalue of 3.2), corresponding to 5.2% of the unexplained variance. The secondary factor was sufficiently dominated by the Rasch dimension to justify the assumption of unidimensionality (Linacre, 2010).

Third, as a check on the internal structure of the measure, we evaluated whether our *a priori* predictions about the relative difficulty of items were correct (Wilson, 2005). Recall that when creating the assessment, we selected items to tap knowledge at each of the four levels on our construct map. The hypothesized difficulty level for each item correlated highly with the empirically derived item difficulty, $r(62) = .85$ $p < .01$. We also used an item-respondent display called a Wright map to help evaluate our difficulty predictions (Wilson, 2005). The Wright map allows for quick visual inspection of whether our construct map correctly predicted relative item difficulties (Figure 2). In brief, a Wright map consists of two columns, one for respondents and one for items. On the left column are *respondents* (i.e., participants). Those with the highest ability scores on the construct are located near the top of the map, while those with the lowest scores are located near the bottom. Assessment *items* are located on the right column. The most difficult items are located near the top of the map and the least difficult ones are near the bottom. The vertical line between these two columns indicates the scale for both the ability and difficulty parameter estimates measured in logits (i.e., log-odds units). The average of the item distribution was set to 0 logits; negative scores indicate items that were easier than average, and positive scores indicate items that were harder than average. The advantage of the logit scale is that it can be used to calculate the probability that a participant of a given ability level will be successful on an item of a particular difficulty. The Wright map shown in Figure 2 is condensed to represent the selected items discussed below. The full assessment and corresponding Wright maps are available from the authors upon request.

As can be seen on the Wright map, the items we had categorized as Level 4 items were indeed the most difficult items (i.e., they clustered near the top with difficulty scores greater than 0); the items we had categorized as Levels 1 and 2 items were indeed fairly easy items (i.e., clustered near the bottom with difficulty scores less than -1); and Level 3 items fell in between. Overall, the distribution of items on the Wright map supports our hypothesized levels of knowledge, progressing in difficulty from a rigid operational view at Level 1 to a comparative relational view at Level 4. After confirming that items were clustered as expected, we added horizontal lines on the Wright map corresponding to approximate cut points between levels. We added these lines to aid discussion, but it should be remembered that the construct is at root a continuous measure and that speaking in terms of levels is merely a conceptual convention.

We also found that student ability levels behaved as expected. First, the ability level of individual students was highly correlated with grade level $r(224) = .72, p < .01$. Second, the correlation between the TCAP math scores for Grade 3 to 6 students and their ability estimates was also high $r(170) = .70, p < .01$. This positive correlation between our assessment and a general standardized math assessment provides some evidence of convergent validity.

The current results, collected from a new population in a different school district, replicate our original findings of adequate psychometric properties for our assessments (Authors, in press). These findings provide strong evidence for the reliability and validity of our mathematical equivalence instrument. In support of our first hypothesis, the results indicate that items from different classes can be measured on a single scale, with a hierarchy of item difficulty that matches our construct map.

Elaborating the Variability in Student Knowledge.

A second goal was to add resolution to our conception of the variability in students' understandings of mathematical equivalence. To illustrate this contribution, we will first discuss how the current model augments the discussions put forth by studies that focus on individual classes of items. We exploit the fact that individual student ability scores, as estimated by the Rasch model, can be used to find specific probabilities that a given student will be successful on a given item. Specifically, we can calculate the probability of any participant's success on any given item using the equation $\Pr(success) = \frac{1}{1+e^{-(\theta - d)}}$ where θ is a participant's ability estimate, and d is the item difficulty estimate. This is a powerful analytical tool, because it allows us to take a single measure (a student's ability score) and use it to predict the types of items with which a student is likely to struggle – without the usual need for resource intensive item-by-item error analysis. The contrasts between these probability estimates for different students will be used to help provide a detailed picture of the variability in student knowledge of mathematical equivalence. We selected 6 representative student ability estimates – corresponding to high and low scores within Levels 2, 3 and 4 – for the purposes of illustration. Table 3 lists the probabilities that students at each of these various ability levels will generate correct answers for selected items.

Open-equation solving items. Several researchers have suggested that students' difficulties with solving open equations should be dependent upon the format of those open equations (Falkner et al., 1999; Weaver, 1973), and the levels of our construct map are based in part upon this idea. The more nonstandard or unfamiliar the format, the more difficult the item should be. Unfortunately, to date no one has tried to quantify the differences in difficulty among nonstandard equations of different formats.

```
PERSON - MAP - ITEM

     PERSONS — LOGITS - ITEMS

          7
          |
          |
          |
          6
          |
        . |
          |
        # 5T
          |
         T|
        . |
          4
    .#### |
          |  If 76+45=121,does 76+45-9=121-9?
      .## |
        # 3  Judge “89+44=87+46” as T/F
     .### |  If 56+85=141,does 56+85-7=141-7?          n+n+n+2=17. Find the value of n.
        # |S
  ####### S|
    .#### 2
          |  c+c+4=16. Find the value of c
  .###### |  What does the equals sign mean?
  ####### |
  ####### 1
    .#### |
       .# |
     .### |
    #### MOM
      ### |  7+6+4=7+□
     #### |  □+2=6+4              Judge “7+6=6+6+1” as T/F
      .## | 13=n+5. Find the value of n.
      .# -1  Judge “8=8” as T/F     Rate “The equal sign means two amounts are the same” as good or bad
       ## |
    ##### |
     .### |
      ### -2
   .#### S|
        # |S
 .####### |  8=6+ □               Judge “4=4+0” as T/F
      .## -3
     #### |
      ### |
        # |
       ## -4
          |
        # |
         T|
         -5T
          | □ +5=9
        . |
        . |
         -6

        EACH "#" IS 2 PEOPLE. EACH "." IS 1 PERSON.
```

Figure 2. The Wright Map. Each “#” represents two respondents, and each “.” represents one respondent. M = Mean, S= Standard Deviation, and T = Two Standard Deviations

The simplest nonstandard open equation items was □ + 5 = 9. This was expected to be the case because this item still adhered to standard operations-equals-answer format, and was classified at Level 1. As shown in Table 3, even the lowest performers were able to solve this item correctly over 90% of the time.

The next most difficult items were those with all operations on the right of the equals sign. Although students on the upper half of the ability continuum showed complete mastery for these items, students of lesser ability were substantially less likely to solve these items than items in operations-equals-answer format. For instance, low level 2 students were expected to solve □ + 5 = 9 correctly 92% of the time, but were expected to solve 8 = 6 + □ correctly only 50% of the time (an odds ratio of 11.5)!

Finally, items became even more difficult when they involved operations on both sides of the equals sign. For instance, □ + 2 = 6 + 4 was considerably more difficult than □ + 5 = 9 for all but the most skilled students. Correct performance for low level 2 students was expected to drop to 10% for this item. Although low level 3 students had largely mastered solving equations with operations on the right, they were only expected to get this item correct 38% of the time. Interestingly, the somewhat longer item, 7 + 6 + 4 = 7 + □ was of nearly identical difficulty despite the fact that the item involves an extra addend and an extra addition sign. This suggests that our construct map is correct in positing the relevant placement of operations as a factor affecting difficulty and leaving the number of addends and whether the unknown is on the left or right side as relatively unimportant factors. In summary, as specified on the construct map, equations with all operations on the right were harder than equations with all operations on the left (the distinction between Level 1 and Level 2), and equations with operations on both sides were harder than equations with operations only on the right (the distinction between Level 2 and

Table 3

Probability of Success on Selected Items by Student Ability Estimate

		STUDENT ABILITY LEVEL (θ = Rasch Ability Estimate)					
		Low Level-2 θ = -2.7	High Level-2 θ = -1.35	Low Level-3 θ = -0.95	High Level-3 θ = 0.10	Low Level-4 θ = 0.61	High Level-4 θ = 3.75
ITEM	Item Difficulty Estimate						
Equation-Solving Items							
□ + 5 = 9	-5.21	0.92	0.98	0.99	1.00	1.00	1.00
8 = 6 + □	-2.71	0.50	0.80	0.85	0.94	0.97	1.00
□ + 2 = 6 + 4	-0.47	0.10	0.29	0.38	0.64	0.75	0.99
7 + 6 + 4 = 7 + □	-0.31	0.08	0.26	0.35	0.60	0.72	0.98
Equation-Structure Items							
4 = 4 + 0 T or F	-2.56	0.47	0.77	0.83	0.93	0.96	1.00
8 = 8 T or F	-1.18	0.18	0.46	0.56	0.78	0.86	0.99
7 + 6 = 6 + 6 + 1 T or F	-0.7	0.12	0.34	0.44	0.69	0.79	0.99
Equals-Sign Items							
Rate “The equal sign means two amounts are the same” as good or bad	-.96	0.15	0.40	0.50	0.74	0.83	0.99
What does the equals sign (=) mean?	1.51	0.01	0.05	0.08	0.20	0.29	0.90
Literal Variable Items							
13 = n + 5	-.75	0.12	0.35	0.45	0.70	0.80	0.99
c + c + 4 = 16	1.76	0.01	0.04	0.06	0.16	0.24	0.88
m + m + m = m + 12	2.67	0.00	0.02	0.03	0.07	0.11	0.75
Properties of Equivalence Items							

Explain if “67 + 86 = 68 + 85” is True or False without computing.	2.15	0.01	0.03	0.04	0.11	0.18	0.83
Explain why, without subtracting the 9, if 76 + 45 = 121 is true, 76 + 45 - 9 = 121 – 9 is true too (could refer them to table 2)	3.48	0.00	0.01	0.01	0.03	0.05	0.57

Level 3).

We examined the textbook series used at the participating schools (Harcourt) to explore whether frequency of exposure to these different equation types parallels improving competence with different equation formats. As expected, in second grade, equations in operations-equals-answer format predominated (86% of instances of the equal sign). By fifth grade, this format was much less frequent (50%), and equations with no operations (21%) or operations on the right (16%) were fairly frequent. However, equations with operations on both sides of the equal sign were infrequent across grade levels (ranging from 0% to 8% of instances of the equal sign). This suggests that simple frequency of exposure does not independently explain increasing success with equations with operations on both sides.

Equation-structure items. We confirmed that acceptance of nonstandard forms was an indicator of sophistication with equivalence knowledge and that certain equation formats are much more difficult to evaluate than others. Behr (1980) suggested that problems should increase in difficulty as they vary in form from equations with operation on the right side to equations with no operations to equations with operations on both sides of the equals sign. Indeed, equations with operations limited to the right side, such as 4 = 4 + 0, though non-standard, were not as difficult for children to grasp as non-canonical equations with no operators (e.g., 8=8) or those with operations on both sides, such as 7 + 6 = 6 + 6 + 1 (see Table 3). High Level 2 students were largely expected to accept 4 = 4 + 0 as true (78% correct), but fall to chance for 8 = 8 (46% correct) and to rarely accept 7 + 6 = 6 + 6 + 1 as true (28% correct). A high Level 3 performer exhibits clear mastery for 4 = 4 + 0 (94%) and is correct the vast majority of the time for 8 = 8 (80%). Performance falls off a bit, however, for the item with operators on both sides of the equals sign (64%).

All told, the data show that although it is correct that whether or not students accept nonstandard equation formats is an important indicator of equivalence knowledge, it further pays to consider *which* nonstandard formats they accept. Some nonstandard formats are much more difficult than others. Moreover, it appears that equation format affects difficulty similarly for both open-equation and equation-structure item types.

Equal-sign items. We were also interested in how students' abilities to give a relational definition of the equals sign was related to their success on other items designed to tap their explicit knowledge of the equals sign, as well as items on open-equation and equation-structure items. First, we should note that requiring that children to provide a relational definition of the equal sign was more difficult than other from the class of *equals-sign* items. Recognizing a relational definition of the equals sign from a list – as opposed to generating one – was much easier. For instance, asking children to rate the phrase, "the equal sign means two amounts are the same" as good or bad was much easier (see Table 3).

Providing a relational definition of the equals sign proved to be very difficult for many children. For example, both high level 3 students (at the beginning relational level) and high Level 2 students (at the flexible operational level) were unlikely to offer a relational definition of the equals sign (20% and 1% probabilities respectively, see Table 3). Similar performance on this item, however, stands in stark contrast to the differences in performance these same students are expected to have on other items. For instance, when asked to evaluate the equation 4 = 4 + 0 as true or false, the high Level 3 student is expected to be successful 93% of the time, whereas the Low Level 2 student is expected to be successful only 47% of the time. Similarly, when asked to solve □ + 2 = 6 + 4, high Level 3 students are successful 64% of the time, despite their general failure to generate a relational definition of the equals sign. By contrast the low Level 2 students are expected to succeed on this item only 10% of the time. By including these diverse measures in one hierarchy, our assessment helps map variability that might otherwise remain uncharted.

Our model also adds resolution when considering those who succeed at offering a relational definition of the equals sign. This is despite the fact that there is far less variability within our sample for those who offer a relational definition than for those who do not. Only 10% of our sample is located a full logit above the difficulty level for this question (i.e., expected to provide a relational definition at least 73% of the time) whereas 59% of our sample is located a full logit below this item on the Wright map (i.e., expected to provide a relational definition less than 27% of the time). This compressed variability on the high end suggests that this item is indeed a good one to pick if searching for a single indicator of mastery. In order to show how our model improves resolution for the high end students, the following discussion will diverge a bit in form, comparing students within a given ability level (high level 4 students) as opposed to comparing students between ability levels.

In our sample, there were 22 students showing 'mastery' for this item, (i.e. whose ability estimates were a full logit above the item difficulty, as defined above). Of these master definers, seven students scored a full two logits higher than the difficulty of the item (i.e., they were expected to get the item correct 88% of the time). The differences between the abilities of the two subgroups of master definers were quite substantial, which becomes apparent when we consider their expected performance on a still more difficult item. Take for example, an item that asks them to provide an explanation using explicit arithmetic properties of equivalence, *76 + 45 = 121 is true. Is 76 + 45 – 9 = 121 – 9 true or false?* (see below for a more detailed discussion of the item). The more highly skilled subgroup of master definers was expected to be successful on this high difficulty item roughly twice as often as the less skilled subgroup (27% vs. 51%). Thus, there is even clearly detectible variability within the mastery level students.

As a final point on students' verbal interpretations of the equals sign, assessing whether or not a child offered a relational definition was not the same as if we were much stricter and coded whether or not a child offered *only* a relational definition (sometimes students offered alternative operational interpretations alongside their relational offerings). We did not include this stricter coding in our Rasch model due to the model's requirements for item independence, but we do have performance information for this stricter criterion. Whereas 26% of our sample could offer a relational definition, only 10% of our sample – or roughly 2 out of every 5 relational definers – provided only relational definitions. Recall from our construct map that we hypothesized that the relational definition could coexist with other less sophisticated definitions. This also accords well with previous findings that different definitions emerge in different contexts (McNeil & Alibali, 2005b; Seo & Ginsburg, 2003).

Equations with Literals. There were three items with literals as variables on each form of the assessment, and the equations varied in whether there were multiple instances of the variable and in whether there were operations on one or both sides of the equation. Using literals extends the range of number sentences by introducing repeated instances of the unknown, something that cannot be easily tested without the use of literals (Carpenter et al., 2003; Jacobs et al., 2007).

The first thing to note is that each of the three items was well aligned with the construct, both according to indicators from classical test theory and measures specific to the Rasch model. Item total correlations were above .45, and infit and outfit statistics of model fit were within acceptable ranges of .5 to 1.5 for each of the items. This is an important point: the items using literals fit the Rasch model as well as other items typically used to study mathematical equivalence. This suggests that this subset of algebraic items loads very heavily on equivalence knowledge.

The next thing to note is that these items (as expected) were not of equal difficulty. Instead, difficulty varied as the format of the item changed. The item $13 = n + 5$ received a difficulty rating of -.75, making it a mid Level 3 difficulty item. It appears that the use of the literal made this item more difficult than other open-equation items with similar format and no literal variable. For example, $8 = 6 + \square$ was significantly easier, with a difficulty of -2.71 (Table 3). Even though use of the literal *n* is logically equivalent to the use of the blank for the missing numeral, it appears that the use of the literal renders the item more difficult. This is perhaps because the literal is less familiar.

The item $c + c + 4 = 16$, on the other hand, was a low Level 4 item. Even though all operations were on the left, this item proved more difficult than the most difficult open equation solving items on the assessment (i.e., $43 + \square = 48 + 76$, difficulty = 1.08). This is presumably because the item involved multiple instances of the unknown, demanding more novel application of the concept of equivalence (Carpenter et al., 2003).

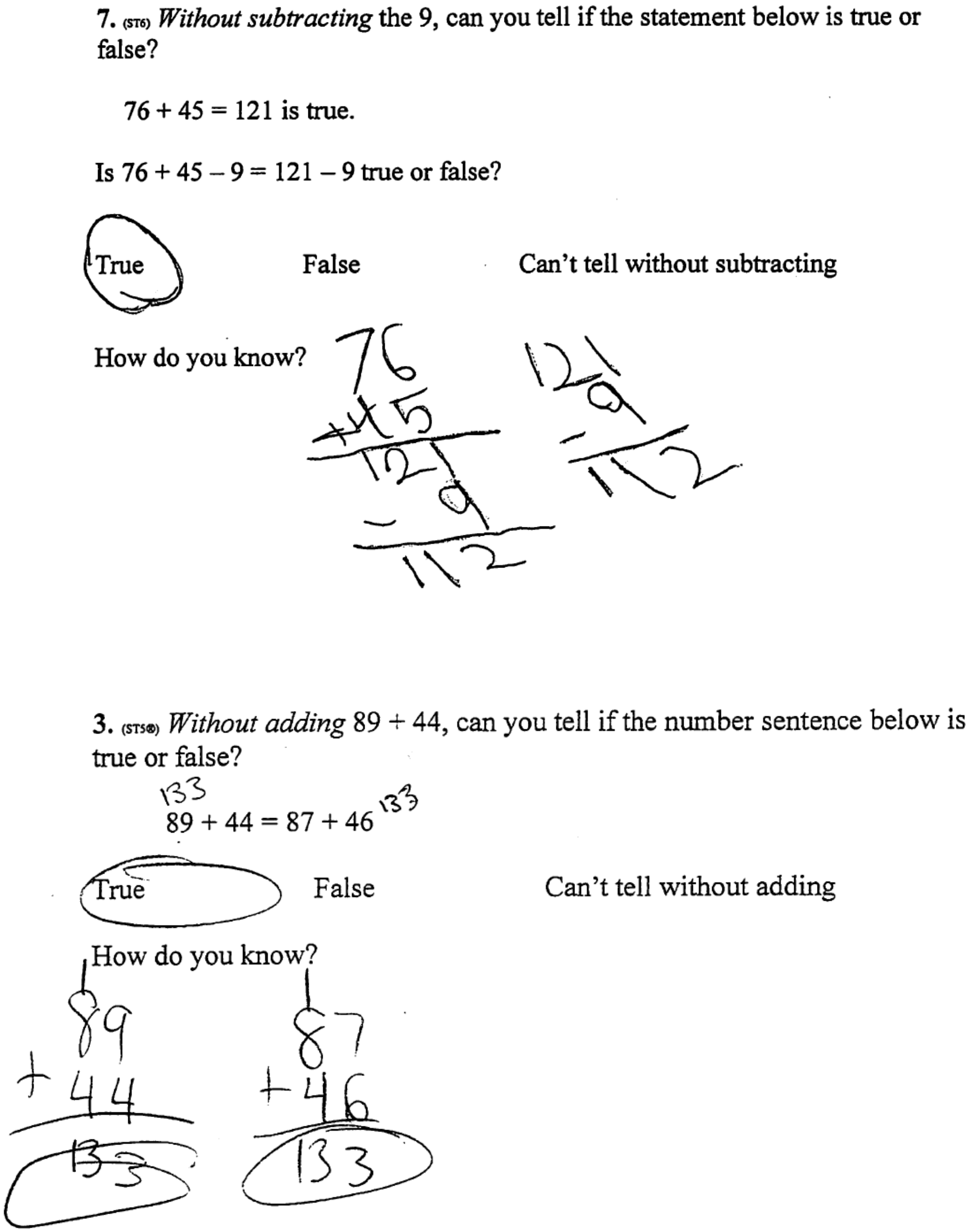
Finally, $m + m + m = m + 12$ emerged as a high difficulty level 4 item (difficulty = 2.67). This item involved both multiple instances of the literal variable and operations on both sides of the equals sign. In fact, it was the most difficult of all assessment items not requiring an explanation. Only 11% of low Level 4 students were expected to solve this item properly, whereas 72% of them are expected to solve $7 + 6 + 4 = 7 + \square$ correctly (see Table 3). It was of even higher difficulty than the item asking for a verbal definition of the equals sign.

In summary, items involving literals: a) fit the model of the equivalence construct well, and b) were more difficult than similarly formatted items, but children with higher knowledge of equivalence seemed to be able use that knowledge to perform better on these less familiar problems. Future iterations of our construct map should include the use of literals as a factor that increases item difficulty.

Use of properties of equivalence. Requiring that students explain the reasoning behind their solutions in terms of the relation between numerical expressions is also thought to explicitly elicit algebraic thought. Although traditional algorithms are based on fundamental properties of equivalence, children rarely receive practice making these properties explicit (Jacobs et al., 2007; Steinberg et al., 1991). These items were intended to go beyond testing procedural proficiency in order to assess whether or not children could express generalizations about properties of equivalence in natural language. Appreciation of these properties allows students to think about numerical expression as something more than a series of calculations(Carpenter et al., 2003).

These algebraic thought items were well aligned with the construct. All but one had infit and outfit statistics below 1.5 and all but one had an item total correlation above .2. Just as with literals, items testing for explicit verbal knowledge of properties of equivalence fit the Rasch model as well as items more typically used to study mathematical equivalence.

As can be seen on the Wright map (Figure 2), these items were the most difficult items on the assessment. To illustrate this point, consider the two items in Figure 3. Even high Level 4 students, who were expected to define the equals sign relationally 90% of the time, are expected to answer these items correctly only 73% and 83% of the time, respectively (Table 3). As predicted by our construct map, it appears that articulating the properties of equivalence represents an advance in equivalence knowledge over interpreting the equals sign in a relational manner. These items help to distinguish between children who all successfully provided relational definitions of the equal sign.



7. (ST6) ***Without subtracting*** **the 9, can you tell if the statement below is true or false?**

76 + 45 = 121 is true.

Is 76 + 45 – 9 = 121 – 9 true or false?

[✓ True] **False** **Can't tell without subtracting**

How do you know?

$$\begin{array}{r} 76 \\ +45 \\ \hline 121 \\ -9 \\ \hline 112 \end{array} \qquad \begin{array}{r} 121 \\ -9 \\ \hline 112 \end{array}$$

3. (ST5®) *Without adding* 89 + 44, can you tell if the number sentence below is true or false?

133 133

89 + 44 = 87 + 46

[✓ True] False Can't tell without adding

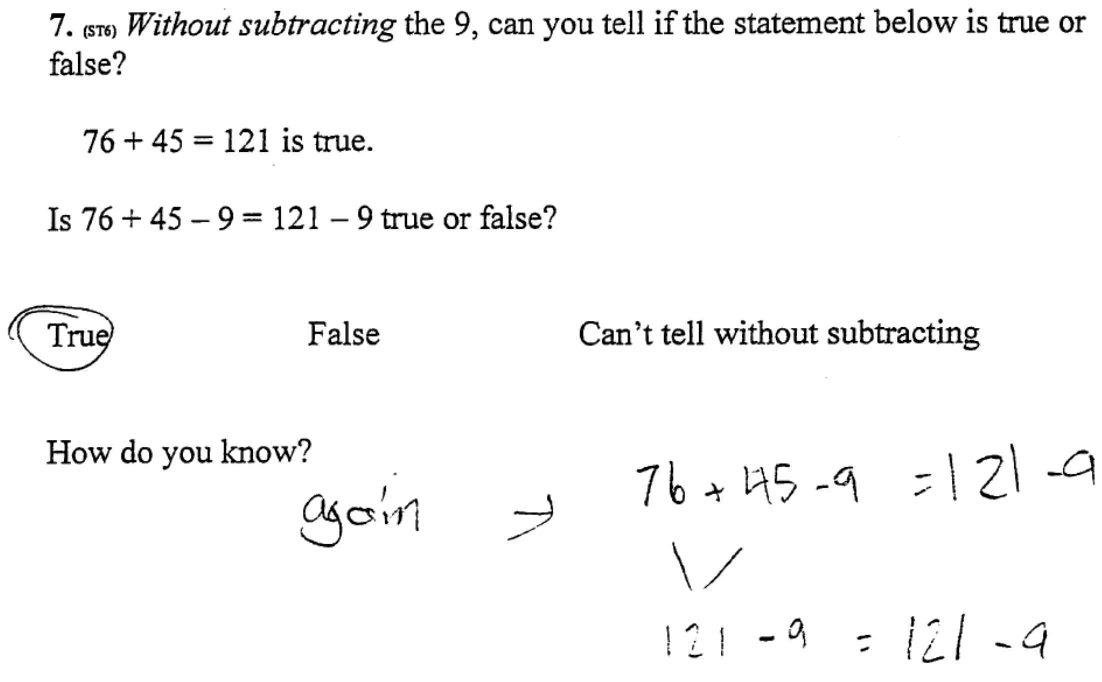
How do you know?

$$\begin{array}{r} \overset{1}{8}9 \\ +44 \\ \hline 133 \end{array} \qquad \begin{array}{r} \overset{1}{8}7 \\ +46 \\ \hline 133 \end{array}$$

Figure 3. Student use of the inefficient solve and compare strategy

Consider the explanations that students gave to these items. Some students could evaluate the truth of these equations via procedural routes, but nonetheless could *not* explain the logical shortcuts that proceed from relational thinking (and thus were scored as 0's). For instance, the examples from two students in Figure 3 correctly state that the expressions are equal but used a resource intensive solve-and-compare strategy to justify their answers. Although the solve-and-compare strategy does require that students realize on some level that the equals sign expresses the interchangeability of each side of an equation, this strategy is not very efficient and does not capture comparative relational thinking (Level 4).

The inefficiency of this strategy stands in relief to the more advanced strategies that use the properties of equivalence (see Table 2). For instance, the student in Figure 4 has generalized



7. (ST6) *Without subtracting* the 9, can you tell if the statement below is true or false?

76 + 45 = 121 is true.

Is 76 + 45 – 9 = 121 – 9 true or false?

[✓ True] False Can't tell without subtracting

How do you know?

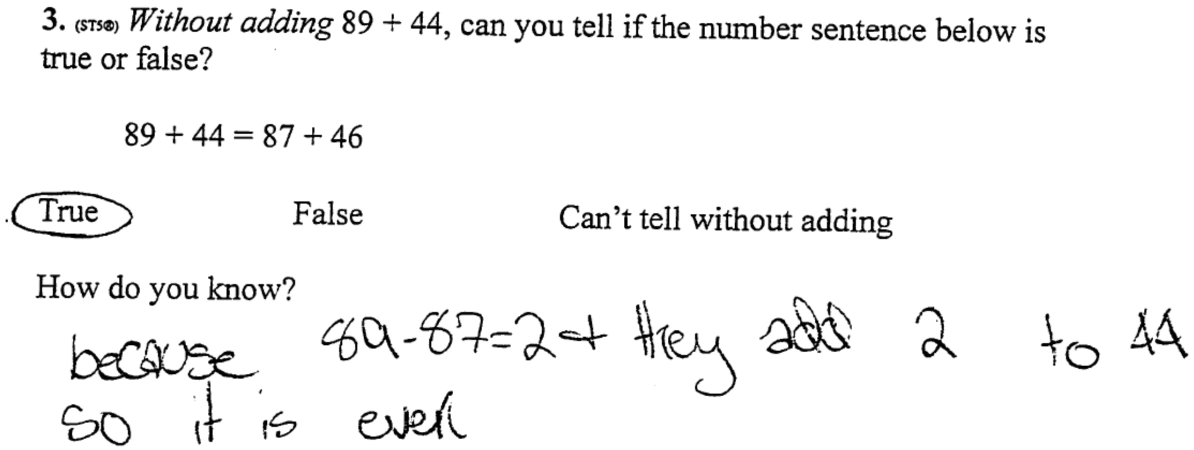
again → 76 + 45 -9 = 121 -9
\/
121 -9 = 121 -9

Figure 4. Student recognition that performing the same operation on both side preserves equivalence

the fact that performing the same operations on both sides of an equation conserves equivalence, a requisite skill for recognizing the link between arithmetic and algebra (Jacobs et al., 2007).

Another example of such advanced equivalence knowledge is reflected by answers that demonstrate use of compensation strategies as in Figure 5 below. In this example, the student has used the fact that 87 is two less than 89, which means that the addend coupled with the 87 must be two more than the one coupled with 89 in order to preserve equality. Carpenter et al, (2003) has argued that such use of compensation strategies instead of full calculation is an indicator of relational thinking.

In summary, items that require explaining properties of equivalence – which are typically thought to assess algebraic thought – fit our model of mathematical equivalence as laid out in our construct map. These items requiring that students explain the properties of equivalence are amongst the most difficult items on the assessment, which helps illustrate an important feature



3. (ST5®) *Without adding* 89 + 44, can you tell if the number sentence below is true or false?

89 + 44 = 87 + 46

[✓ True] False Can't tell without adding

How do you know?

because 89-87=2+ they add 2 to 44
so it is even

Figure 5. Student use of compensation strategy

of our method: it can be used to explore new potential items that measure knowledge of mathematical equivalence. Kieran (1981) has previously suggested that failures on some more advanced mathematical items may in fact stem from incorrect conceptions of equivalence. As more candidate items are generated, our method can assess a) whether or not these new potential items measure the construct of mathematical equivalence knowledge, and if so, b) where they fall in the hierarchy of difficulty.

DISCUSSION

Several decades of research have catalogued the difficulties that American elementary school children have with understanding the concept of mathematical equivalence (e.g., Alibali, 1999; Baroody & Ginsburg, 1983; Behr, 1980; Falkner et al., 1999; Jacobs et al., 2007; Li et al., 2008; Lindvall & Ibarra, 1980; Matthews & Rittle-Johnson, 2009; McNeil, 2007; Powell & Fuchs, 2010; Weaver, 1973). The current study adds resolution to that picture by reconciling previously incommensurable measurement items onto a single scale. We replicated and extended the results from a previous study to develop this measure. Our findings reaffirmed that these diverse items could be integrated and provided further support for the validity of our measure and of our construct map of equivalence knowledge. Moreover, because we used a new sample, our findings provide early evidence for the generalizability of our construct map. Below, we discuss how our measure helps expand our abilities to measure students' knowledge of mathematical equivalence, both in terms of variability and in terms of the construct's link to typical algebra items. We then discuss some of its current limitations and reflect upon ways that it might be improved in the future.

Resolving variability in student knowledge

Our construct map and measure augments the information gleaned from any single class of equivalence items. First, not all nonstandard equation formats are equally challenging. As detailed in our construct map, the more that an equation varies from the standard a + b = c format, the more difficult it is likely to be. In particular, difficulty seems primarily to depend upon the location of the operations in the equation. Difficulty increases as form changes from all operations on the left side, to all on the right side, and becomes most difficult when operations are included on both sides of the equals sign. Although the comparative difficulties of these problems have been suggested in several studies (Baroody & Ginsburg, 1983; Carpenter et al., 2003; Weaver, 1973) differences in performance according to problem format has rarely been quantified. Our findings confirm the importance of equation format by offering a wide range of problem formats and testing the effect of problem format on both open equation-solving items and equation structure items. They also indicate that differences in item difficulty are quite substantial.

Second, we were able to compare the difficulties of different classes of items all thought to tap understanding of equivalence. Focusing across rows in a given column in Table 3 allows us to compare the differences in the probabilities of success on different items for students of a given ability level. The table helps highlight an interesting point about how equation format affects difficulty in a similar fashion across classes – it appears that open-equation and equation structure items were of similar difficulty when they were in the same equation formats (e.g., 8 = 6 + □ vs. 4 = 4 + 0 T/F). The table also highlights some interesting differences. For instance, consider low Level 4 students. Table 3 clearly reveals that providing a relational definition of the equals sign is typically much more difficult than working with equations with operations on both sides, be it solving open-equation or accepting nonstandard formats as valid. The ability to accept and solve equations in nonstandard formats and the ability to give a relational definition of the equals sign lie substantially far apart on the scale of increasing equivalence knowledge.

This point illustrates a more general feature of our model in that it helps make clear exactly how wide the variability is among students who might otherwise be labeled as being at the same ability level. For example, even though 75% of our sample failed to define the equals sign relationally, we could detect differences within this group because of the nature of our measurement scale.

Additional Empirical Evidence For The Link Between Algebra And Math Equivalence

This is the first instrument that explicitly maps typical algebraic items on to more basic items measuring mathematical equivalence. In so doing, we showed that algebraic items involving literal symbols load heavily on the construct of equivalence (e.g., $m + m + m = m + 12$). We saw that students with higher equivalence knowledge were more likely to solve equations involving literals. Importantly, this effect for equivalence knowledge prevailed even though these children presumably had no more experience with literal variables than their peers in the same classrooms. These findings corroborate the claims of others linking young children's equivalence knowledge to algebraic thinking (Knuth et al., 2006; Lindvall & Ibarra, 1980; MacGregor & Stacey, 1997; Steinberg et al., 1991).

Analysis of performance on equations involving literal symbols suggests two trends: First, the use of literals may have added difficulty relative to the use of blanks, even though they were logically equivalent. Moreover, it seems that equation format seems to have influenced difficulty with equations involving literals in much the same way that it influenced difficulty for equations that involved no literals. Namely, equations with operations on both sides were harder for children to solve than those that involved operators on a single side only. The use of multiple instances of an unknown, which is only possible with literals, also increased item difficulty.

Our method also allowed us to examine student's explicit use of the properties of equivalence in order to demonstrate comparative relational thought. These items were among the most difficult on our instrument, surpassing even equations with literals. That is, providing verbal explanations for the properties of equivalence on equations using only numbers was more demanding – some times much more so – than was solving equations using relatively unfamiliar literal variables. This highlights the fact that procedural competence with mathematical equivalence can sometimes precede ability to articulate rules governing the domain (Karmiloff-Smith, 1986).

Limitations

Our measure and construct map have been developed with two different samples of elementary school students. In both samples, schools were using traditional math curriculum that did not focus on mathematical equivalence. Children's developing knowledge of equivalence is largely thought to be due to exposure – practice with standard operations-equals-answer formats encourages the development of operational viewpoints (Alibali et. al., 2007; Li, et. al., 2008; McNeil & Alibali, 2005a, Seo & Ginsburg, 2003), Children exposed to multiple formats from early on infrequently develop such operational patterns, demonstrating more facility with the concept of equivalence earlier (Li et al., 2008). Thus, our construct map may only apply to children from similar educational backgrounds. For example, we would expect the construct map to look different for a population of Chinese elementary school students or for students using a curriculum focused on varied equation structures (e.g., the Wynroth curriculum studied by

Baroody & Ginsburg, 1983). With more wide scale testing, we can investigate the extent to which our construct map generalizes across different student populations.

A second limitation of the current study is that it relies on students' written responses on a paper-and-pencil assessment. Students likely have knowledge that is not revealed on paper-and-pencil tests. For example, Jacobs et al (2007) used a student interview to reveal whether students used relational thinking to solve some of the problems on the written test, including a prompt to encourage use of relational thinking. We incorporated these prompts for relational thinking in our written assessment, but children often will say more than they will write. Integrating items from a structured interview with the written assessment would help reveal the impact of assessment format on student performance.

The Power Of The Method

Our final point is about method. Our intent was to build upon prior research in order to gain more leverage from the items typically used to measure equivalence knowledge. Our contribution, therefore, is first and foremost one of method. Science has historically been constrained by the limits of method (Kuhn, 1996). We have presented evidence that suggests that our current science regarding math equivalence has similarly been limited by method. The use of diverse items without a unifying metric does not allow us to take advantage of the full leverage that an integrated instrument affords.

Research in mathematics education is pregnant with potential for practical uses of Wilson's construct-modeling approach. In our case, we chose to use the method due to the lack of an integrated assessment of equivalence knowledge. In another case, Clements, Sarama, and Liu (2008) used the approach to construct an assessment for measuring mathematics ability in 3-5 year old children. The authors found that existing measures such as the Woodcock-Johnson III had not been validated for children in this young age range. They consulted experts, developed a construct map, and designed items to measure the construct. In the final analysis, the construct-based approach holds potential to help fill lacunae in the field wherever measures are wanting or non-existent.

In summary, we demonstrated how our new assessment built using a construct modeling approach stands to enrich our knowledge of children's understanding of mathematical equivalence. Our new method can: a) compare across item classes, b) increase resolution as far as item difficulty, even within classes, and c) allow for conceptual expansion of items that might demand equivalence knowledge. Moreover, it may eventually allow us to expand the difficulty level upward to find places where more advanced students and adults falter in their understanding/activation of equivalence knowledge (see Kieran, 1981; MacGregor & Stacey, 1997; McNeil & Alibali, 2005a).

References

Adelman, C. (2006). The Toolbox Revisited: Paths to Degree Completion From High School Through College. *US Department of Education*, 223.

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

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*(3), 221-247.

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

Baroody AJ, Lai ML, Mix KS. The development of number and operation sense in early childhood. In: Saracho O, Spodek B, eds. *Handbook of research on the education of young children*. Mahwah, NJ: Erlbaum; 2006: 187-221.

Behr, M. (1980). How Children View the Equals Sign. *Mathematics Teaching*, *92*, 13-15.

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

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking Mathematically: Integrating Arithmetic and Algebra in Elementary School.* Heinemann, 361 Hanover Street, Portsmouth, NH 03801-3912 (Paperback: $24.50). Web site: www. heinemann. com.

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

Cobb, P. (1987). An investigation of young children's academic arithmetic contexts. *Educational Studies in Mathematics*, *18*(2), 109-124.

De Corte, E., & Verschaffel, L. (1981). Children's solution processes in elementary arithmetic problems: Analysis and improvement. *Journal of Educational Psychology*, *73*(6), 765-779.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children’s understanding of equality: A foundation for algebra. *Teaching Children Mathematics*, *6*(4), 232-236.

Ginsburg, H. (1977). *Children's arithmetic: The learning process*. New York: D. Van Nostrand Co.

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education, 40*(3), 241-250.

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*(3), 258.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data* 1. *Cognition*, *23*(2), 95-147.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics*, *12*(3), 317-326.

Kieran, C. (1992). The learning and teaching of school algebra. *Handbook of research on mathematics teaching and learning*, 390-419.

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

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*(4), 297.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer Verlag.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. University of Chicago press Chicago.

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*(2), 195-217.

Linacre, J. M. (2010). A user's guide to winsteps rasch-model computer programs Available from http://www.winsteps.com/winman/index.htm?copyright.htm

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*, 50-62.

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), 1-21.

McNeil, N. M. (2008). Limitations to teaching children 2 + 2 = 4: Typical arithmetic problems can hinder learning of mathematical equivalence. *Child Development, 79*, 1524-1537.

McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental psychology*, *43*(3), 687-694.

McNeil, N. M., & Alibali, M. W. (2005a). Knowledge change as a function of mathematics experience: All contexts are not created equal. *Journal of Cognition and Development*, *6*(2), 285-306.

McNeil, N. M., & Alibali, M. W. (2005b). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*(4), 883-899.

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*(3), 367-385.

Moses, R. P., & Cobb, C. E. (2001). *Radical equations: Math literacy and civil rights*. Beacon Pr.

National Research Council. (1998). *The nature and role of algebra in the K-14 curriculum*. Washington, DC: National Academy Press.

National Science Board. (2002). *Science and engineering indicators—2002* (NSB-02–1). Arlington, VA: National Science Foundation.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development*, *6*(4), 449-468.

Perry, M., Church, R.B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. Cognitive Development, 3(4), 359–400.

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*(2), 381-394.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* MESA Press, 5835 S. Kimbark Ave., Chicago, IL 60637; e-mail: MESA@ uchicago. edu; web address:

www. rasch. org; telephone: 773-702-1596 fax: 773-834-0326 ($20).

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1-15.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other?. *Journal of educational psychology*, *91*(1), 175-189.

Seo, K. H., & Ginsburg, H. P. (2003). "You've got to carefully read the math sentence...": Classroom context and children's interpretations of the equals sign. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Sherman, J., & Bisanz, J. (2009). Equivalence in symbolic and nonsymbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology, 101*(1), 88-100.

Siegler, R. S. (1998). *Emerging minds: The process of change in children's thinking*. Oxford University Press, USA.

Steinberg, R. M., Sleeman, D. H., & Ktorza, D. (1991). Algebra students' knowledge of equivalence of equations. *Journal for research in mathematics education*, 112-121.

Warren, E. (n.d.). Young children’s understanding of equals: A longitudinal study. *INTERNATIONAL GROUP FOR THE PSYCHOLOGY OF MATHEMATICS EDUCATION.*

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*(1), 45-56.

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 97-116.