tapraid5/zcz-edu/zcz-edu/zcz00111/zcz2293d11z | xppws | S=1 | 9/23/10 | 3:25 | Art: 2009-0604

Journal of Educational Psychology

0022-0663/10/$12.00 DOI: 10.1037/a0021334

Assessing Knowledge of Mathematical Equivalence: A Construct-Modeling Approach

Bethany Rittle-Johnson and Percival G. Matthews
Vanderbilt University

Roger S. Taylor
State University of New York at Oswego

Katherine L. McEldoon
Vanderbilt University

Knowledge of mathematical equivalence, the principle that 2 sides of an equation represent the same value, is a foundational concept in algebra, and this knowledge develops throughout elementary and middle school. Using a construct-modeling approach, we developed an assessment of equivalence knowledge. Second through sixth graders (N = 175) completed the assessment on 2 occasions, 2 weeks apart. Evidence supported the reliability and validity of the assessment along a number of dimensions, and the relative difficulty of items was consistent with the predictions from our construct map. By Grade 5, most students held a basic relational view of equivalence and were beginning to compare the 2 sides of an equation. This study provides insights into the order in which students typically learn different aspects of equivalence knowledge. It also illustrates a powerful but underutilized approach to measurement development that is particularly useful for developing measures meant to detect changes in knowledge over time or after intervention.

Keywords: algebra, mathematical equivalence, measurement development, construct map

The widespread goal of “algebra for all” underscores the importance of making algebra accessible to all students, not just those who aspire to careers in math and science. For example, high-school students who completed Algebra II were five times more likely to graduate from college than those who completed only Algebra I (Adelman, 2006). There is an emerging consensus that, to increase students’ success in algebra, educators must reconceptualize the nature of algebra as a continuous strand of reasoning throughout school rather than a course saved for middle or high school (National Council of Teachers of Mathematics, 2000). Part of this effort entails assessing children’s early algebraic thinking.

In the current paper, we describe development of an assessment of one component of early algebraic thinking: knowledge of mathematical equivalence. Mathematical equivalence, typically represented by the equal symbol, is the principle that two sides of an equation represent the same value. We employed a construct-modeling approach (Wilson, 2003, 2005) and developed a construct map (i.e., a proposed continuum of knowledge progression) for students’ knowledge of mathematical equivalence. We used the construct map to develop a comprehensive assessment, administered the assessment to students in Grades 2 to 6, and then used the data to evaluate and revise the construct map and the assessment. The findings provide insights into the typical sequence in which learners acquire equivalence knowledge. The study also illustrates an approach to developing measures that are particularly useful for detecting changes in knowledge over time or after intervention.

Need for Reliable and Valid Measures

Too often, researchers in education and psychology use measures that have not gone through a rigorous measurement development process, a process that is needed to provide evidence for the validity of the measures (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 1999). For example, Hill and Shih (2009) found that less than 20% of studies published in the *Journal for Research in Mathematics Education* over the past 10 years had reported on the validity of the measures. As they noted,

Bethany Rittle-Johnson, Percival G. Matthews, and Katherine L. McEldoon, Department of Psychology and Human Development, Peabody College, Vanderbilt University; Roger S. Taylor, Department of Psychology, State University of New York at Oswego.

This research was supported with funding from National Science Foundation Grant DRL-0746565 to Bethany Rittle-Johnson, as well as Institute of Education Sciences, U.S. Department of Education, training grants R305B040110 and R305B080008. The opinions expressed are those of the authors and do not represent the views of the funding agencies. A special thanks to the students and teachers at St. Edward School for participating in this research. Thanks to Holly Harris, Robin Hill, Caitlyn Majeika, and Jadie Pinkston 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 No. 0552, Nashville, TN 37203. E-mail: b.rittle-johnson@vanderbilt.edu

> Without conducting and reporting validation work on key independent and dependent variables, we cannot know the extent to which our instruments tap what they claim to. And without this knowledge, we cannot assess the validity of inferences drawn from studies. The AERA/APA/NCME (testing) standards heavily emphasize the collection and reporting of such information in research studies. (Hill & Shih, 2009, p. 248)

The lack of evidence for the reliability and validity of measures applies to previous measures of mathematical equivalence knowledge. First, there is no standard measure of equivalence knowledge; rather, researchers use their own self-designed measures (e.g., Baroody & Ginsburg, 1983; Jacobs, Franke, Carpenter, Levi, & Battey, 2007; Kieran, 1981; Li, Ding, Capraro, & Capraro, 2008; Rittle-Johnson, 2006). Second, we could not find a study that reported evidence for the validity of a particular measure. Third, only two studies reported information on the reliability of a measure, and this was restricted to reporting Cronbach's alpha on scales containing about four items (Jacobs et al., 2007; Li et al., 2008).

These measurement issues may help to explain some discrepancies in past findings. For example, Knuth, Stephens, McNeil, and Alibali (2006) reported that a majority of middle-school students in their study did not understand equivalence, whereas Matthews and Rittle-Johnson (2009) found that a majority of fifth graders in their sample did. Knuth et al. relied on students' written definition of the equal sign; Matthews and Rittle-Johnson relied on students' ability to solve equations with mathematical operations on both sides of the equation (e.g., $3 + 7 + 5 = 3 + \square$). Although both approaches measure equivalence knowledge, providing a verbal definition is likely more difficult than solving problems, and differences in how knowledge of equivalence is assessed can lead to seemingly contradictory claims that may be resolved with closer attention to measurement.

Knowledge of Mathematical Equivalence

Although few previous studies have paid careful attention to measurement issues, a large number of studies have assessed children's knowledge of mathematical equivalence (sometimes called mathematical equality). It is a fundamental algebraic concept that is accessible in the elementary grades (e.g., Jacobs et al., 2007; McNeil, 2007). Understanding mathematical equivalence requires understanding that the values on either side of the equal sign are the same. This specific knowledge about mathematical equations is distinct from knowledge of numerical equivalence. By 4 years of age, children can match sets of objects on the basis of quantity, suggesting that they have knowledge of numerical equivalence (e.g., Gelman & Gallistel, 1986; Mix, 1999). Unfortunately, students do not seem to link their knowledge of numerical equivalence for sets of objects to interpreting and solving written equations like $8 + 4 = \square + 5$ (Falkner, Levi, & Carpenter, 1999; Sherman & Bisanz, 2009).

Knowledge of mathematical equivalence is a critical prerequisite for understanding higher level algebra (MacGregor & Stacey, 1997). In particular, it is necessary for competently performing the same operation on both sides of an equation and for understanding equivalent expressions (Kieran, 1992; Steinberg, Sleeman, & Ktorza, 1990). For example, middle-school students who correctly define the equal sign are much more likely than those who do not to solve equations correctly (Knuth et al., 2006).

Given the importance of mathematical equivalence, it is of concern that students often fail to understand this concept. Many view the equal sign *operationally*, as a command to carry out arithmetic operations, rather than *relationally,* as an indicator of equivalence (e.g., Jacobs et al., 2007; Kieran, 1981; McNeil & Alibali, 2005b). Evidence for this has primarily come from three different classes of equivalence tasks: (a) *equation-solving items*, such as $8 + 4 = \square + 5$; (b) *equation-structure items*, such as deciding if $3 + 5 = 5 + 3$ is true or false; and (c) *equal-sign-definition* items. To solve equations such as $8 + 4 = \square + 5$, most elementary-school students either add the numbers before the equal sign or add all the given numbers (e.g., respond that the answer is 12 or 17; Falkner et al., 1999). Indeed, in a broad range of studies spanning 35 years of research, a majority of first through sixth graders treated the equal sign operationally when solving equations with operations on the right side or both sides of an equation, sometimes with only 10% of students solving the equations correctly (e.g., Alibali, 1999; Behr, Erlwanger, & Nichols, 1980; 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).

Similarly, students tend to not be comfortable with equation structures without a standard $a + b = c$ structure (e.g., *operations-equals-answer* structure). When asked to evaluate whether equations are true or false, most elementary-school children indicated that only equations with an operations-equals-answer structure are true (Baroody & Ginsburg, 1983; Behr et al., 1980; Falkner et al., 1999; Freiman & Lee, 2004; Li et al., 2008; Molina & Ambrose, 2006; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003). For example, in informal interviews, 6- and 7-year-olds rejected nonstandard equation structures and rewrote them into an operations-equals-answer structure, such as rewriting $3 = 3$ as $0 + 3 = 3$ (Behr et al., 1980).

Finally, when defining the equal sign, first and second graders typically define it operationally as "what it adds up to" or "when two numbers are added, that's what it turns out to be" (Behr et al., 1980; Ginsburg, 1977; Seo & Ginsburg, 2003). Students' responses are not much more sophisticated in the later grades, with almost half of middle-school students in two recent studies giving operational definitions of the equal sign (Alibali, Knuth, Hattikudur, McNeil, & Stephens, 2007; Knuth et al., 2006).

Performance on all three classes of items for tapping children's developing knowledge of equivalence suggests that an operational understanding of equivalence develops as the default knowledge representation and is not easy to overcome. However, difficulty understanding equivalence is not universal, as it is not prevalent in elementary-school students educated in some other countries, such as China and Taiwan (Li et al., 2008; Watchorn, Lai, & Bisanz, 2009).

The primary source of the difficulty that U.S. children have in understanding mathematical equivalence is thought to be their prior experiences with the equal sign (e.g., Baroody & Ginsburg, 1983; Carpenter, Franke, & Levi, 2003; Falkner et al., 1999; McNeil, 2007, 2008). Elementary-school children are thought to receive little direct, explicit instruction on the meaning of the equal sign. Rather, students may infer an incorrect meaning of the equal sign from repeated experience with limited equation structures (e.g., Baroody & Ginsburg, 1983; Carpenter et al., 2003; Falkner et al., 1999; McNeil, 2007, 2008). An analysis of two second-grade textbooks identified very few instances in which the equal sign was not presented in an operations-equals-answer structure (Seo & Ginsburg, 2003). Falkner et al. (1999) speculated that

> not much variety is evident in how the equals sign is typically used in the elementary school. Usually, the equals sign comes at the end of an equation and only one number comes after it. With number sentences, such as $4 + 6 = 10$ or $67 - 13 = 54$, the children are correct to think of the equals sign as a signal to compute. (p. 232)

This operational understanding of equivalence is difficult to overcome. For example, second- and third-grade children received direct instruction that the equal sign meant "the same as" during an experimental intervention. If this instruction was presented in the context of equations with an operations-equals-answer structure, they continued to solve equations with operations on both sides incorrectly (McNeil, 2008).

What is less clear is how a correct understanding of mathematical equivalence develops. Recent research has shown that a substantial minority (often around 30%) of students in elementary school can solve equations with operations on both sides of the equal sign correctly, particularly fourth and fifth graders (e.g., Freiman & Lee, 2004; Matthews & Rittle-Johnson, 2009; McNeil, 2007; McNeil & Alibali, 2004, 2005b; Oksuz, 2007; Rittle-Johnson, 2006). By the end of middle school, a majority gave a relational definition of the equal sign (e.g., 60% of students in Alibali et al., 2007). McNeil (2007) noted that topics that contradict an operational view of the equal sign, such as equivalent fractions, inequalities, and pre-algebra, are discussed in later elementary grades, and proposed this should help weaken an operational view and strengthen a relational view of the equal sign. By sixth grade, students are being exposed to equations in a variety of formats in their textbooks (e.g., with no operations, such as 12 in = 1 foot, $2/4 = 1/2$, and $x = 4$; McNeil et al., 2006). It is unclear when such variability in problem structures is introduced in textbooks, as analyses of how the equal sign is presented in third- to fifth-grade textbooks have not been reported. Overall, correct understanding of equivalence may develop through implicit processes and may take many years to develop.

Construct Map for Mathematical Equivalence

Our primary goal in the current study was to develop an assessment that could detect systematic changes in children's knowledge of equivalence across elementary-school grades (second through sixth). To accomplish this, we utilized Mark Wilson's construct-modeling approach to measurement development (Wilson, 2003, 2005). The core idea is to develop and test a *construct map*, which is a representation of the continuum of knowledge through which people are thought to progress for the target construct. This continuum is often broken into different levels to help conceptualize the knowledge progression, but it is important to note that the continuous nature of the model means that the levels should not be interpreted as discrete stages.

Our construct map for mathematical equivalence is presented in Table 1, with less sophisticated knowledge represented at the T1 bottom and more advanced knowledge represented at the top. The four knowledge levels differ primarily in the types of equations with which students are successful, starting with equations in an operations-equals-answer structure, then incorporating equations with operations on the right or no operations, and finally incorporating equations with operations on both sides (initially with single-digit numbers and eventually with multidigit numbers that increase the value of using more sophisticated strategies). Past research suggests the structure of the equation should be a primary influence on performance, regardless of item class (e.g., solving an equation vs. evaluating an equation as true or false), although this prediction has not been explicitly tested. Prior research also indicates that knowing a relational definition of the equal sign is related to success on equations with operations on both sides of the equal sign (Alibali et al., 2007; Rittle-Johnson & Alibali, 1999).

Past research has focused on two levels, a rigid operational view (Level 1) and a basic relational view (Level 3; see Table 1). We hypothesized that there would be a transition phase between these two views, labeled *Level 2: Flexible operational view*. In particular, we predicted that students would become less rigid and would successfully solve equations and evaluate and encode equation structures that are atypical but remain compatible with an operational view of the equal sign, such as equations that are "backwards" (e.g., __ = 2 + 5; Behr et al., 1980) or that contain no

Table 1
Construct Map for Mathematical Equivalence Knowledge

Level	Description	Core equation structures
Level 4: Comparative relational	Successfully solve and evaluate equations by comparing the expressions on the two sides of the equal sign, including using compensatory strategies and recognizing that performing the same operations on both sides maintains equivalence. Recognize relational definition of equal sign as the best definition.	Operations on both sides with multidigit numbers or multiple instances of a variable
Level 3: Basic relational	Successfully solve, evaluate, and encode equation structures with operations on both sides of the equal sign. Recognize *and generate* a relational definition of the equal 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 equal 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 equal sign operationally.	Operations on left: a + b = c (including when blank is before the equal sign)

Note. Italics indicate ideas that may need to be revised, based on the current data.

operations (e.g., $3 = 3$). By second grade, students have moderate levels of success solving these types of equations (Freiman & Lee, 2004; Weaver, 1973) and accepting statements with no operations as true (Seo & Ginsburg, 2003). In addition, Carpenter et al. (2003) proposed using equations in these formats to help transition students to understanding equations with operations on both sides. We did not expect children at this level to define the equal sign relationally.

We also hypothesized that some elementary-school students would be developing knowledge of equivalence that went beyond a basic relational understanding. This *Level 4: Comparative Relational* thinking captures success solving equations and evaluating equation structures by comparing the expressions on the two sides of the equal sign. As a result, the students' reasoning need not be tied to specific computations. For example, students with a comparative understanding know that doing the same things to both sides of an equation maintains its equivalence, without needing to verify the equivalence relation with full computation (e.g., "If $56 + 85 = 141$, does $56 + 85 - 7 = 141 - 7$?"; Alibali et al., 2007; Steinberg et al., 1990). They also use compensatory strategies to ease calculations with large numbers, such as quickly solving $28 + 32 = 27 + □$ by recognizing that 27 is 1 less than 28, so the unknown must be 1 more than 32 (Carpenter et al., 2003). We also expected that a relational definition of the equal sign would be dominant at this level, with students considering a relational definition of the equal sign to be the best definition, and that students would have an explicit awareness that the equal sign divides the equation into two sides (Rittle-Johnson & Alibali, 1999).

Although the construct map is presented as having four levels for descriptive purposes, our conception of the construct, as well as our statistical model, is continuous. 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).

Current Study

We used our construct map to guide creation of an assessment of mathematical equivalence knowledge, with items chosen to tap knowledge at each level of the construct map with a variety of item classes. We administered an initial long version of the assessment to children in Grades 2 through 6, and 2 weeks later we administered a revised, shorter version of the assessment. We used an item response model to evaluate our construct map in addition to using classical test theory methods to provide additional evidence for the reliability and validity of the assessment (e.g., internal consistency, test–retest reliability). To provide some insights into a potential source of knowledge change, we also analyzed the textbooks used at the participating school for frequency of presentation of different equation structures.

Method

Participants

Second- through sixth-grade students from 10 classrooms at an urban parochial school participated. There were 184 participants with parental consent who completed the initial assessment, but 9 of these students were absent when we administered the revised assessment. Of the 175 students who completed the revised assessment, 37 were in second grade (17 girls, mean age = 7.7 years), 43 were in third grade (27 girls, mean age = 8.9 years), 33 were in fourth grade (14 girls, mean age = 9.8 years), 34 were in fifth grade (17 girls, mean age = 10.7 years), and 28 were in sixth grade (13 girls, mean age = 11.7). The students were from a working- to middle-class community, and approximately 20% of students in the participating grades were from minority groups (approximately 8% African American and 5% Hispanic).

The school used the Iowa Tests of Basic Skills (ITBS; see http://www.education.uiowa.edu/itp/itbs/) as a standardized measure of educational progress. Students' percentile ranks and grade equivalent scores in math and reading on the ITBS were obtained from student records. On average, students scored in the 60th percentile in math (range = 4th to 99th percentile) and the 67th percentile in reading (range = 14th to 99th percentile).

Each teacher completed a brief survey on how much time her students had spent on five activities related to equivalence during the current school year, using a 4-point scale ranging from none to a week or more. The most common activity was comparing numbers, and many students had also spent a week or more solving or seeing equations without an operations-equals-answer structure (see Table 2). The second graders had spent a fair amount of time discussing the meaning of the equal sign, but the older students had not. Finally, some second and third graders had solved equations with literal variables, and all the fourth- to sixth-grade students had.

Test Development Procedure

Overview. We developed a pool of possible assessment items from past research on mathematical equivalence. Items were selected and modified so that each level of the construct map was covered by at least two items in each of the three common item classes identified in the literature review: solving equations, evaluating the structure of equations, and defining the equal sign. We piloted potential items with 24 second- to fourth-grade students at a local afterschool program (servicing a different school) to screen out inappropriate items. We worked with these pilot students one-on-one, and, on the basis of their responses and input, we eliminated or reworded confusing items and created an initial assessment instrument that could be administered within a single 45-min class period. This initial assessment was administered to 10 classes of second- to sixth-grade students. Analyses of these results and input of a domain expert informed the creation of two shorter, comparable forms of a revised assessment, which were administered to the same students 2 weeks later. All versions of the assessment had three sections based on the three item classes.

Equation-solving items. These items tapped students' abilities to solve equations at the four knowledge levels and were taken from four previous studies (Carpenter et al., 2003; Jacobs et al., 2007; Matthews & Rittle-Johnson, 2009; Weaver, 1973). The Level 4 items were adapted from the work of Carpenter and colleagues. For example, to solve $67 + 84 = □ + 83$, students can compare the expressions and know they need to add 1 to 67, because 83 is 1 less than 84, and answer 68. Students were encouraged to "try to find a shortcut so you don't have to do all the

Table 2
Teacher Responses to "How Much Time Have Students Spent on the Following Activities This School Year?" by Grade

Activity	Grade 2 Teacher A	Grade 2 Teacher B	Grade 3 Teacher A	Grade 3 Teacher B	Grade 4	Grade 5	Grade 6
Solving problems in which the equal sign is not at the end (e.g., 4 + _ = 9; 3 + 6 = _ + 8)	Week+	3–5 days	None	3–5 days	1–2 days	3–5 days	3–5 days
Seeing problems in which the equal sign is not at the end (e.g., 8 = 8; 5 + 2 = 2 + 5)	3–5 days	3–5 days	None	Week+	Week+	3–5 days	Week+
Discussing meaning of the equal sign	Week+	3–5 days	1–2 days	1–2 days	1–2 days	1–2 days	1–2 days
Solving equations with variables (e.g., 4 + 7 = t and t + 8 = 14)	None	3–5 days	3–5 days	None	Week+	3–5 days	Week+
Comparing numbers using >, <, =	Week+	Week+	Week+	Week+	Week+	3–5 days	Week+

Note. For Grades 4–6, one teacher at each grade level taught math.

adding," and sometimes pressure was used to discourage full computation. At this level, they can also compare sides to simplify and solve equations with multiple instances of a variable. For instance, given the equation $n + n + n + 2 = 17$, they can solve it by first recognizing that $n + n + n$ must equal 15 and then using the fact that three 5s are 15 to solve the problem (Jacobs et al., 2007). The initial assessment had 28 equation-solving items, and the revised assessment had 11.

Equation-structure items. These items were designed to probe students' knowledge of valid equation structures, and the equations varied according to the criteria outlined in Table 1. A majority of items asked students to evaluate equations as true or false, sometimes with follow-up prompts to explain their evaluations, and were taken from four previous studies (Baroody & Ginsburg, 1983; Behr et al., 1980; Carpenter et al., 2003; Warren, 2003). Other items asked students to reconstruct equations from memory (to measure encoding of the equation structure) or to identify the two sides of an equation (Matthews & Rittle-Johnson, 2009; McNeil & Alibali, 2004; Rittle-Johnson & Alibali, 1999). The most advanced items assessed whether students would (a) compare the expressions on either side of the equal sign to determine whether an equation such as 89 + 44 = 87 + 46 was true (e.g., explain "true, because 89 is 2 more than 87, but 44 is 2 less than 46") or (b) accept doing the same thing to both sides of an equation, based on items from three studies (Alibali et al., 2007; Carpenter et al., 2003; Steinberg et al., 1990). The initial assessment had 31 equation-structure items; the revised assessment had 18.

Equal-sign items. These items were designed to probe students' explicit knowledge of the equal sign. A core item asked students to define the equal sign (e.g., Behr et al., 1980; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003). Students were also asked to rate definitions of the equal sign (McNeil & Alibali, 2005a; Rittle-Johnson & Alibali, 1999) and to select the best definition of the equal sign; these questions were inspired by methods used in the psychology literature to assess people's knowledge of concepts (Murphy, 2002). Two easier items probed if students could recognize that the equal sign can be used to indicate equivalent values when no operators are involved (Sherman & Bisanz, 2009) and could recognize the equivalence of symbolic expressions (e.g., 5 + 5 is equal to 6 + 4; Rittle-Johnson & Alibali, 1999). The initial assessment had 13 equal-sign items; the revised assessment had eight.

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 1 of the correct answer to allow for minor calculation errors. For the five explanation items, students received a point if they mentioned the equivalent relation between values on the two sides of the equation; see Appendix A for scoring details for individual items.

Revised assessment. We revised the initial assessment in order to have two shorter, comparable forms of the assessment. Items on the initial assessment were eliminated if (a) they had poor psychometric properties on more than one of three key indices described in the item screening section and a domain expert had identified them as inappropriate (5 items) or (b) they were redundant with other items that had stronger psychometric properties (7 items). In addition, we revised two items that had been flagged on multiple indices. Based on accuracy data and the suggestion of the domain expert, the two pairs of items with the equation structure $\square + b = c$ or $a + \square = c$ were reclassified from Level 2 to Level 1 (2SOL and 4STR in Table 3). T3

We created two comparable, short versions of the assessment (37 items each) based upon these revisions. Items from the initial assessment were paired on the basis of three criteria: level of equivalence tapped, question stem, and proportion correct on the initial assessment. Whenever possible, we used one item from each pair on Form 1 and the other item on Form 2. However, sometimes we needed to create a new, similar version of an item (5 items) or use the same item on both assessments (4 items). We used this step-by-step item-matching procedure to ensure that content would be comparable across forms, as content similarity is a prerequisite for meaningful score-equating procedures (Kolen & Brennan, 2004). The number of items of each class was based on the number of available items in the literature and the time requirements for completion of the different item classes. The items from one form of the revised assessment are presented in Appendix A.

Test administration. Assessments were administered on a whole-class basis. At Time 2, both forms were randomly distributed in each classroom. A spiraling procedure was used to ensure random equivalence of the groups responding to each form (i.e., Form 1 was given to the first student and Form 2 to the next, with alternation thereafter). A member of the research team read the directions aloud for each section and used a preplanned script to help answer any questions participants raised. The team member also enforced a time limit for each section (see Appendix A) in

Table 3
Item Statistics for Mathematical Equivalence Assessment

Level	Form	Item name[a]	Item summary	Expert rating	Accuracy	Item total correlation	Item difficulty	*SE* of difficulty	*t* test between pairs	Infit MSQ	Outfit MSQ
1	1	1DEF.L1.1	Identify pair that is equal to 3 + 6	2.75	0.76	0.56	−1.43	0.31	2.81*	0.93	0.99
	2	1DEF.L1.2	Identify pair that is equal to 6 + 4		0.87	0.52	−2.83	0.39		0.80	0.30
	1	2SOL.L1.1	□ + 5 = 9	2.75	0.98	0.20	−4.66	0.74	N/A[b]	1.00	0.31
	2	2SOL.L1.2	4 + □ = 8		1.00	NA	N/A	0.00		1.00	1.00
	1	3STR.L1.1	Judge "8 = 5 + 10" as false	3.50	0.82	0.52	−1.95	0.33	0.63	0.96	0.52
	2	3STR.L1.2	Judge "5 + 5 = 5 + 6" as false		0.77	0.69	−1.65	0.34		0.75	0.35
	1	4STR.L1.1	Recall □ + 2 = 5	2.50	0.94	0.08	−3.61	0.49	0.15	1.22	2.49
	2	4STR.L1.2	Recall □ + 2 = 5		0.92	0.27	−3.51	0.44		1.31	1.17
2	1	5DEF.L2.1	2 nickels < > 1 dime. Select choice that shows they are the same.	2.50	0.75	0.27	−1.34	0.31	0.98	1.55	3.81
	2	5DEF.L2.2	5 pennies < > 1 nickel. Select choice that shows they are the same.		0.78	0.58	−1.79	0.34		1.00	0.74
	1	6SOL.L2.1	7 = □ + 3	3.75	0.92	0.37	−3.18	0.43	0.60	0.92	0.39
	2	6SOL.L2.2	8 = 6 + □		0.87	0.42	−2.83	0.39		0.95	2.34
	1	7STR.L2.1	Judge "8 = 8" as true or false	4.25	0.80	0.50	−1.73	0.32	1.91*	1.01	0.84
	2	7STR.L2.2	Judge "3 = 3" as true or false		0.69	0.63	−0.85	0.33		1.11	1.73
	1	8STR.L2.1	Judge "8 = 5 + 3" as true or false	4.75	0.84	0.57	−2.17	0.34	0.53	0.69	0.32
	2	8STR.L2.2	Judge "7 = 3 + 4" as true or false		0.79	0.63	−1.91	0.35		0.78	0.71
	1	9STR.L2.1	Explain the judgment of 8STR.L2.1	4.75	0.72	0.56	−1.05	0.31	0.85	0.95	1.29
	2	9STR.L2.2	Explain the judgment of 8STR.L2.2		0.75	0.56	−1.44	0.34		1.14	1.04
3	1	10DEF.L3.1	What does the equal sign mean?	4.25	0.43	0.63	1.18	0.30	1.60	1.04	0.95
	2	10DEF.L3.2	What does the equal sign mean?		0.37	0.61	1.86	0.30		0.94	1.50
	1	11DEF.L3.1	"The equal sign means two amounts are the same." Is this a good or not good definition?	3.75	0.84	0.50	−2.17	0.34	4.13*	0.85	0.42
	2	11DEF.L3.2	"The equal sign means the same as." Is this a good or not good definition?		0.62	0.59	−0.24	0.32		1.34	1.98
	1	12SOL.L3.1	5 + □ = 6 + 2	4.33	0.61	0.79	−0.24	0.30	0.46	0.58	0.46
	2	12SOL.L3.2	□ + 2 = 6 + 4		0.62	0.83	−0.24	0.32		0.61	0.38
	1	13SOL.L3.1	3 + 6 = 8 + □	5.00	0.60	0.83	−0.15	0.30	0.21	0.50	0.37
	2	13SOL.L3.2	3 + 4 = □ + 5		0.60	0.75	−0.04	0.32		0.83	0.75

(*table continues*)

Table 3 (*continued*)

Level	Form	Item name[a]	Item summary	Expert rating	Accuracy	Item total correlation	Item difficulty	*SE* of difficulty	*t* test between pairs	Infit MSQ	Outfit MSQ
3	1	14SOL.L3.1	□ + 9 = 8 + 5 + 9	4.50	0.44	0.76	1.10	0.30	1.72	0.73	0.51
	2	14SOL.L3.2	8 + □ = 8 + 6 + 4		0.55	0.87	0.36	0.31		0.53	0.37
	1	15SOL.L3.1	4 + 5 + 8 = □ + 8	5.00	0.52	0.78	0.48	0.30	2.09*	0.66	0.69
	2	15SOL.L3.2	7 + 6 + 4 = 7 + □		0.64	0.79	−0.45	0.33		0.77	0.59
	1	16SOL.L3.1	8 + 5 − 3 = 8 + □	4.75	0.51	0.67	0.57	0.30	0.71	0.95	0.95
	2	16SOL.L3.2	6 – 4 + 3 = □ + 3		0.56	0.72	0.26	0.32		0.88	0.97
	1	17STR.L3.1	Judge “4 + 1 = 2 + 3” as true or false	4.75	0.63	0.80	−0.32	0.30	0.54	0.54	0.41
	2	17STR.L3.2	Judge “6 + 4 = 5 + 5” as true or false		0.66	0.85	−0.56	0.33		0.49	0.29
	1	18STR.L3.1	Judge “3 + 1 = 1 + 1 + 2” as true/false	4.50	0.60	0.55	−0.15	0.30	0.13	1.20	1.36
	2	18STR.L3.2	Judge “7 + 6 = 6 + 6 + 1” as true/false		0.60	0.72	−0.09	0.33		0.91	2.62
	1	19STR.L3.1	Explain the judgment of 17STR.L3.1	5.00	0.57	0.82	0.12	0.30	0.82	0.54	0.51
	2	19STR.L3.2	Explain the judgment of 17STR.L3.2		0.62	0.82	−0.24	0.32		0.64	0.39
	1	20STR.L3.1	Judge “5 + 3 = 3 + 5” as true/false	4.50	0.69	0.72	−0.87	0.30	0.54	0.67	0.39
	2	20STR.L3.2	Judge “31 + 16 = 16 + 31” as true/false		0.66	0.76	−0.63	0.33		0.78	0.59
	1	21STR.L3.1	Recall 5 + 2 = □ + 3	2.50	0.86	0.16	−2.42	0.36	0.78	1.50	3.62
	2	21STR.L3.2	Recall 5 + 2 = □ + 3		0.80	0.43	−2.03	0.35		1.33	1.47
	1	22STR.L3.1	Recall □ + 5 = 8 + 7 + 5	2.50	0.48	0.38	0.83	0.30	0.46	1.70	1.81
	2	22STR.L3.2	Recall □ + 5 = 8 + 7 + 5		0.47	0.46	1.03	0.31		1.61	1.89
4	1	23DEF.L4.1	Which definition of the equal sign is the best?	4.25	0.50	0.35	0.65	0.30	3.06*	1.82	2.07
	2	23DEF.L4.2	Which definition of the equal sign is the best?		0.36	0.37	1.95	0.30		1.46	4.52
	1	24SOL.L4.1	□ + 55 = 37 + 54 Try to find a shortcut.	4.50	0.34	0.68	1.90	0.30	1.18	0.81	0.55
	2	24SOL.L4.2	43 + □ = 48 + 76 Try to find a shortcut.		0.43	0.71	1.40	0.30		0.83	0.75
	1	25SOL.L4.1	67 + 84 = □ + 83 Try to find a shortcut.	4.75	0.40	0.76	1.45	0.30	1.39	0.65	0.42
	2	25SOL.L4.2	898 + 13 = 896 + □ Try to find a shortcut.		0.49	0.74	0.85	0.31		0.82	0.63
	1	26SOL.L4.1	c + c + 4 = 16 Find the value of c.	4.25	0.32	0.47	2.08	0.30	2.43*	1.27	1.65
	2	26SOL.L4.2	n + n + n + 2 = 17 Find the value of n.		0.47	0.54	1.03	0.31		1.35	1.95
4	1	27STR.L4.1	Judge “89 + 44 = 87 + 46” as True or False without computing	4.75	0.42	0.65	1.27	0.30	0.12	0.94	0.84
	2	27STR.L4.2	Judge “67 + 86 = 68 + 85” as True or False without computing		0.45	0.62	1.22	0.30		1.09	1.35
	1	28STR.L4.1	Explain the judgment of 27STR.L4.1	5.00	0.17	0.52	3.43	0.35	1.06	0.69	0.38
	2	28STR.L4.2	Explain the judgment of 27STR.L4.2		0.24	0.53	2.92	0.33		0.84	0.65
	1	29STR.L4.1	Identify two sides in 4 + 3 + 6 = 2 + □	2.25	0.24	0.53	2.76	0.32	0.80	0.95	1.22
	2	29STR.L4.2	Identify two sides in 8 + 2 + 3 = 4 + □		0.22	0.41	3.13	0.33		1.16	2.41
	1	30STR.L4.1	If 76 + 45 = 121, does 76 + 45 − 9 = 121 − 9?	5.00	0.24	0.28	2.76	0.32	0.75	1.36	4.16
	2	30STR.L4.2	If 56 + 85 = 141, does 56 + 85 − 7 = 141 − 7?		0.22	0.48	3.11	0.34		0.86	0.83
	1	31STR.L4.1	Explain the judgment of 30STR.L4.1	5.00	0.09	0.37	4.50	0.44	0.10	0.90	0.40
	2	31STR.L4.2	Explain the judgment of 30STR.L4.2		0.10	0.38	4.56	0.45		0.88	0.58

Note. MSQ = mean square.

[a] Labeling convention for item names: Sequential numbering of items, followed by abbreviation for the item class (DEF designates equal-sign-definition items, SOL designates equation-solving items, and STR designates equation-structure items), followed by the Level (L1 for Level 1, etc.), followed by whether it was presented on Form 1 or Form 2. For example, 1DEF.L1.1 is an equal-sign definition item at Level 1 presented on Form 1. [b] A test of equivalence could not be calculated for this pair because all students taking Form 2 completed the item correctly.

* $p < .05$.

order to ensure that students had time to get to all three sections. For second and third graders, a member of the research team also read the directions aloud for each new subset of items to reduce the reading demands of the assessment.

Expert ratings. Expert screening of items on one form of the revised assessment was obtained from four mathematics education researchers who each had over 10 years of experience conducting research on elementary-school children's knowledge of algebra. Each expert rated every item on a scale from 1 to 5 (1 = *not essential*, 3 = *important but not essential*, 5 = *essential*) based on its perceived importance for knowledge of mathematical equivalence. Gathering expert ratings is common practice in measurement development and supports the face validity of the items within a target community (AERA/APA/NCME, 1999).

Measurement Model

We used a Rasch model in addition to methods from classical test theory to evaluate the assessment. Rasch modeling is a one-parameter member of the item response theory (IRT) family (Bond & Fox, 2007). The Rasch model considers both respondent ability and item difficulty simultaneously, estimating the probability that a particular respondent will answer a particular item correctly (Rasch, 1980). A graphical display of the results, known as a Wright map, allowed us to interpret the parameters estimated by our Rasch model in terms of our construct map (Wilson, 2005). We used Winsteps software version 3.68.0.2 to perform all IRT estimation procedures (www.winsteps.com).

The Rasch model estimation procedure also provides information on the goodness of fit between empirical parameter estimates and the measurement model, thus providing indicators of potentially problematic items. In particular, *infit* statistics measure unexpected responses to items with difficulty levels close to respondents' ability estimates. *Outfit* statistics, on the other hand, measure unexpected responses to items with difficulty levels markedly different from respondents' ability estimates. Ideal infit and outfit mean square values are near 1. Values substantially above 1 indicate items that contribute less toward the overall estimate of the latent variable and are most problematic, and values substantially below 1 indicate items that have less variance than expected. Popular criteria favor infit/outfit mean square values that lie between 0.5 and 1.5 (Linacre, 2010).

Item Screening

We screened the 37 items on each form of the revised assessment for sound psychometric properties. Three of the 37 items on each form were Level 1 items meant to check that students were paying attention (e.g., 3 + 4 = □). Accuracy for these items was near 100%, so they were not included in the analyses because they were not diagnostic. We excluded three additional items on each form from further analysis (one equation-structure item and two equal-sign items per form), because there were multiple indicators that they were not good items (i.e., both item-total correlations below .2 and infit and/or outfit mean square values above 1.5). This screening resulted in 31 items on each form with acceptable psychometric properties: 16 equation-structure items, 5 equal-sign items, and 10 equation-solving items. The complete list of items is presented in Table 3.

Results

In presenting our results, we focus on the revised forms of the assessment. Data from the initial assessment are used as supporting evidence when appropriate.

Evidence for Reliability

At the most basic level, assessments must be able to yield reliable measurements. Internal consistency, as assessed by Cronbach's alpha, was high for both of the revised assessments (Form 1 = .94; Form 2 = .95). Performance on the assessment was also very stable between testing times. Test–retest reliability was calculated by computing the correlation between performance on the subset of 28 items that had been given in both the initial and the revised assessment (3 of the items from the revised assessment were not on the initial assessment). There was a high test–retest correlation overall both for Form 1, $r(26) = .94$, and for Form 2, $r(26) = .95$. Finally, the five explanation items on the revised assessments were analyzed for interrater reliability. An independent coder coded responses for 20% of the sample, with a mean exact agreement of 0.99 for Form 1 (range = .96–1.00) and .97 for Form 2 (range = .87–1.00). Overall, both forms of the assessment appeared to yield reliable measures of student performance.

Equating of Scoring Across Forms

We sought to equate scores across forms for two reasons. First, we wanted to establish comparable alternate forms for future intervention or longitudinal research, in which it is helpful to have multiple forms of an assessment. Second, we wanted to be able to evaluate the validity of our construct across all items, instead of separately by form. We used a random groups design within IRT to calibrate the scores from the two forms (Kolen & Brennan, 2004). Item difficulty estimates for both forms of an assessment are calibrated so that the item difficulty of each form is mean centered around zero. As long as groups are equivalent, the ability estimates of participants taking both versions are placed on the same scale, requiring no further transformation or additional equating procedures.

Several indicators confirmed the equivalence of students who completed the two forms. Because a spiraling technique was used to distribute the assessments, the forms were administered to similar numbers of students ($n_{\text{form1}} = 88$, $n_{\text{form2}} = 87$), and the distribution of forms was even within each grade level. Groups were also equivalent in age (mean age for both forms was 9.6 years) and on average ITBS reading grade equivalent scores (Form 1 = 5.75, Form 2 = 5.82) and math grade equivalent scores (Form 1 = 5.45, Form 2 = 5.27).

We also checked to ensure that the two test forms demonstrated similar statistical properties according to classical test theory measures. First, they had virtually identical mean accuracy scores (57% on each form). Second, they had similar mean item discrimination scores (Form 1 = 63%, Form 2 = 69%). Item discrimination scores are an indicator of how well each item discriminates between the top and bottom performers on the assessment and are calculated by finding the difference in percent correct on each item for the top 27% and bottom 27% of students in terms of total score (Rodriguez, 2005). Third, the correlation between mean accuracy on the paired items across the two forms was very high, $r(29) = .94$, $p < .01$.

As a final check on our equating procedure, we compared the estimated item difficulties from the IRT model of paired items across the two forms. Twenty-five of the 31 matched pairs received equivalent item difficulty estimates as indicated by between-samples *t* tests (see Table 3). Differences in accuracy on the remaining six pairs may reflect knowledge and skills not included in our equivalence construct, such as computational fluency. For instance, 1DEF.L1 asks students to evaluate which pair of numbers is equal to the pair in the question stem. The correct answer to the problem on Form 1 is 2 + 7, and the correct answer on Form 2 is 5 + 5. Compared to 2 + 7, problems like 5 + 5 are more often solved by direct retrieval and more rarely solved incorrectly (Ashcraft, 1992). Nevertheless, these six problems still clustered in the appropriate range of item difficulty scores, as predicted by our construct map.

In sum, we confirmed that our forms were administered to equivalent groups, that they demonstrated similar statistical properties, and that similar difficulty estimates were received for most paired items. With these criteria met, it was reasonable to use a random groups design in IRT to calibrate the scores from the two forms, placing all item difficulties and student abilities on the same scale. Hence, the following discussion of validity considers all items simultaneously, placed on the same scale.

Evidence for Validity

Multiple measures provided evidence for the validity of our measure of mathematical equivalence, according to four of the validity categories specified by AERA/APA/NCME (1999).

Evidence based on test content. Experts' ratings of items provided evidence in support of the face validity of the test content. The four experts rated most of the test items to be *important* (rating of 3) to *essential* (rating of 5) items for tapping knowledge of equivalence. The mean validity rating for test content was 4.1 (see Table 3 for the average rating on each item).

Evidence based on internal structure—dimensionality. We conducted several analyses to evaluate 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 principal-components analysis of the residuals 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.

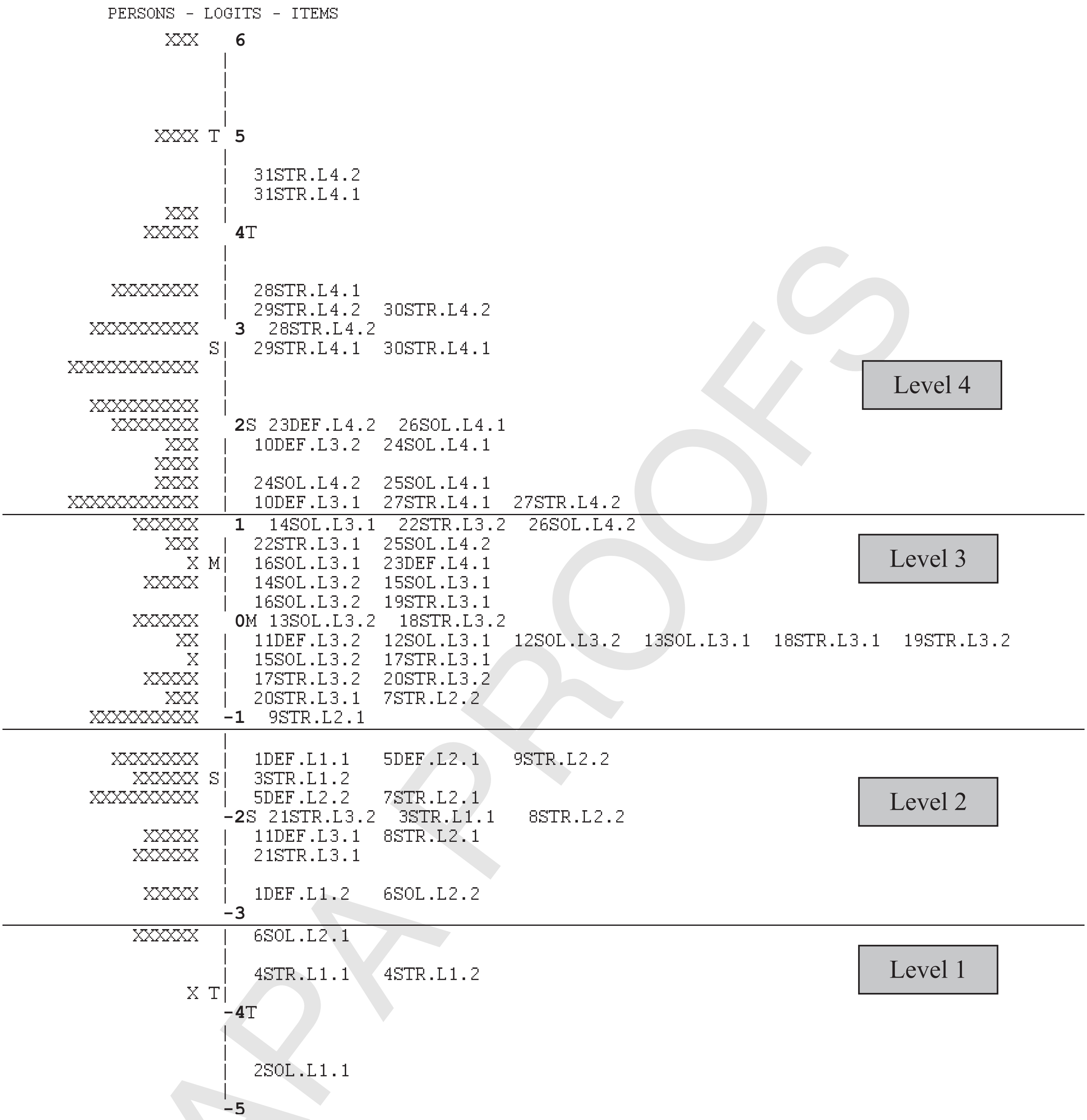
The Rasch model accounted for 57.2% of the variance in the present data set. A principal-components analysis on the residuals indicated that the largest secondary factor accounted for 2.2% of the total variance (eigenvalue = 3.2), corresponding to 5.2% of the unexplained variance. The secondary factor was sufficiently dominated by the Rasch dimension to justify the assumption of unidimensionality (Linacre, 2010).

As an additional check on dimensionality, we conducted a series of confirmatory factor analyses (CFA). We explored three possible factor structures: (a) a one-factor model for all items; (b) a two-factor model, grouping items that have been said to tap knowledge of procedures in past intervention research (equation-solving items) and items that have been said to tap knowledge of concepts (equation-structure and equal-sign definition items; Matthews & Rittle-Johnson, 2009; Rittle-Johnson, 2006; Rittle-Johnson & Alibali, 1999); and (c) a three-factor model, grouping items by the three item classes. We performed the CFAs with Mplus 4.2 (B. Muthén & Muthén, 2006). Because the items were scored dichotomously (wrong-right), the CFA was computed with tetrachoric correlations. Because scores on dichotomous items did not follow a normal distribution, we used the WLSMV estimator, which employs weighted least square estimates with robust standard errors, as recommended by L. K. Muthén (2004). The models had minor problems with some empty cells in between-items correlations, but model estimation terminated normally (see Appendix B for the correlation matrix for the one-factor model). To evaluate the models, we examined the fit indices suggested by Hu and Bentler (1999), namely, the chi-square-based Bentler comparative fit index (CFI) and the residual-based standardized root mean square residual (SRMR), using standards recommended by Tabachnick and Fidell (2007). All models had a very good CFI estimate (CFI = 0.980, 0.980, 0.981 for the one-, two-, and three-factor models, respectively), indicating acceptable fit. According to the residual-based SRMR, however, none of the models showed very good fit (SRMR = 0.121, 0.119, 0.118 for the one-, two-, and three-factor models, respectively; target value ≤ 0.08). Yu (2002) has reported that SRMR does not perform well with binary variables, so we are not as confident in the results of this index. Clearly, the fit of the three different models were very similar, so the increased complexity of a two- or three-factor model did not seem justified.

Overall, a single factor captured a majority of the variance and performance on individual items, suggesting that our construct was unidimensional. The extremely small improvements in fit when additional factors were added, combined with little theoretical justification for additional factors, suggested that including additional factors was not warranted.

Evidence based on internal structure—Wright map. As a second check on internal structure, 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 one of four levels on our construct map. We used an item-respondent map (i.e., a Wright map) generated by the Rasch model to evaluate our construct map. In brief, a Wright map displays participants and items on the same scale. In the left column, respondents (i.e., participants) with the highest estimated ability on the construct are located near the top of the map. In the right column, the items of greatest difficulty are located near the top of the map. The vertical line between the two columns indicates the scale for parameter estimates measured in logits (i.e., log-odds units), which are the natural logarithm of the estimated probability of success divided by the estimated probability of failure on an item. The advantage of using the logit scale is that it results in an equal interval linear scale that is not dependent on the particular items or participants used in estimating the scores. 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 Wright map shown in Figure 1 allows for quick visual inspection of whether our construct map correctly predicted relative item difficulties (see Table 3 for specific item difficulty scores). As can be seen, the items we had categorized as Level 4 F1



```
PERSONS - LOGITS - ITEMS
         XXX   6
              |
              |
              |
              |
        XXXX T 5
              |
              |  31STR.L4.2
              |  31STR.L4.1
         XXX  |
       XXXXX   4T
              |
              |
    XXXXXXXX  |  28STR.L4.1
              |  29STR.L4.2  30STR.L4.2
  XXXXXXXXXX   3  28STR.L4.2
             S|  29STR.L4.1  30STR.L4.1
XXXXXXXXXXXX  |
              |                                                      Level 4
  XXXXXXXXXX  |
    XXXXXXXX   2S 23DEF.L4.2  26SOL.L4.1
         XXX  |  10DEF.L3.2  24SOL.L4.1
        XXXX  |
        XXXX  |  24SOL.L4.2  25SOL.L4.1
XXXXXXXXXXXX  |  10DEF.L3.1  27STR.L4.1  27STR.L4.2
      XXXXXX   1  14SOL.L3.1  22STR.L3.2  26SOL.L4.2
         XXX  |  22STR.L3.1  25SOL.L4.2
           X M|  16SOL.L3.1  23DEF.L4.1                              Level 3
       XXXXX  |  14SOL.L3.2  15SOL.L3.1
              |  16SOL.L3.2  19STR.L3.1
      XXXXXX   0M 13SOL.L3.2  18STR.L3.2
          XX  |  11DEF.L3.2  12SOL.L3.1  12SOL.L3.2  13SOL.L3.1  18STR.L3.1  19STR.L3.2
           X  |  15SOL.L3.2  17STR.L3.1
       XXXXX  |  17STR.L3.2  20STR.L3.2
         XXX  |  20STR.L3.1  7STR.L2.2
  XXXXXXXXXX  -1  9STR.L2.1
              |
    XXXXXXXX  |  1DEF.L1.1   5DEF.L2.1   9STR.L2.2
      XXXXXX S|  3STR.L1.2
  XXXXXXXXXX  |  5DEF.L2.2   7STR.L2.1                               Level 2
              -2S 21STR.L3.2  3STR.L1.1   8STR.L2.2
       XXXXX  |  11DEF.L3.1  8STR.L2.1
      XXXXXX  |  21STR.L3.1
              |
       XXXXX  |  1DEF.L1.2   6SOL.L2.2
              -3
      XXXXXX  |  6SOL.L2.1
              |
              |  4STR.L1.1   4STR.L1.2                               Level 1
           X T|
              -4T
              |
              |
              |  2SOL.L1.1
              |
              -5
```

Figure 1. Wright map for the mathematical equivalence assessment. In the left column, each X represents one person, with least knowledgeable people at the bottom. In the right column, each entry represents an item, with the easiest items at the bottom. The vertical line between the two columns indicates the scale for parameter estimates measured in logits (i.e., log-odds units). Along the vertical line, M indicates the mean, S indicates 1 standard deviation above or below the mean, and T indicates 2 standard deviations above or below the mean. These statistics are included for the participants (i.e., persons; left of center) and for the items (right of center). Refer to Table 3 for details on each item.

items were indeed the most difficult (clustered near the top, with difficulty scores greater than 1), the items we had categorized as Levels 1 and 2 items were indeed fairly easy (clustered near the bottom, with difficulty scores less than −1), and Level 3 items fell in between. Overall, 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. This was confirmed by Spearman's rank order correlation between hypothesized difficulty level and empirically derived item difficulty, $\rho(62) = .84$, $p < .01$.

To evaluate whether individual items were at the expected level of difficulty, we used standard errors to construct confidence intervals around item difficulty estimates. We flagged items that failed to cluster within the empirically derived boundaries of their respective difficulty levels (i.e., Level 4 items with difficulty above 1, Level 3 items with difficulty between 1 and −1, Level 2 items with difficulties between −1 and −3, and Level 1 items with difficulties below −3; see Figure 1). Seven of the 62 items across the two forms of the assessment failed to cluster as expected: 1DEF.L1.1, 3STR.L1.1, 3STR.L1.2, 10DEF.L3.2, 11DEF.L3.1, 21STR.L3.1, and 21STR.L3.2. We briefly consider each of these items in turn.

1DEF.L1.1, 3STR.L1.1, and 3STR.L1.2 were all expected to be Level 1 items but proved more difficult. On the latter two, students needed to identify false equations as false. The equations were in nonstandard formats, and we expected students to easily identify them as false, even if for the wrong reason. The poorer than expected accuracy may indicate general uncertainty elicited by being asked to evaluate a variety of unfamiliar equation structures as true or false. These items were not critical to our construct map and perhaps should be dropped in future iterations. 1DEF.L1.1 may reflect unexpected computational difficulty discussed above, as the parallel item on Form 2 was at the expected level of difficulty.

The remaining mismatched items were all expected to be at Level 3, which suggests that the construct map may need to be refined. 10DEF.L3.2 asked students to provide a definition of the equal sign and was somewhat more difficult than expected. Its twin (10DEF.L3.1) was also difficult, so generating a relational definition may be better classified as a Level 4 item. 11DEF.L3.1 asked students whether “the equal sign means two amounts are the same” is a good definition of the equal sign, and it was easier than expected. Its paired item asked whether “the equal sign means the same as” is a good definition and was at the predicted level of difficulty. It may be that the phrasing “two amounts” provides easier access to the concept of equality. Finally, 21STR.L3.1 and 21STR.L3.2 were identical items on the two forms that asked students to reproduce the equation $5 + 2 = \square + 3$ from memory. These items were considerably easier than expected. This result might be explained by the fact that encoding a problem correctly is necessary for solving it correctly (Siegler, 1976), so successful encoding of a particular structure may precede successful solving of problems with that structure in some circumstances. We will carefully monitor the performance of these items in the future as we continue to validate our assessment and refine the construct map.

The range in difficulty of the items was appropriate for the target population. As shown in Figure 1, the range of item difficulties matched the spread of participant locations quite well (i.e., there were sufficiently easy items for the lowest performing participants and sufficiently difficult items for the highest performing participants). In addition, ability estimates increased with grade level (see Figure 2). As expected, mean ability estimates progressively increased as grade level increased, $\rho(173) = .76$, $p < .01$. F2

Evidence based on relation to other variables. We examined the correlation between students’ standardized math scores on the ITBS and students’ estimated ability on our equivalence assessment (two students were excluded because we did not have ITBS scores for them). As expected, there was a significant positive correlation between scores on the equivalence assessment and grade equivalent scores on the ITBS for mathematics, even after we had partialed out students’ reading scores on the ITBS, $r(86) =$

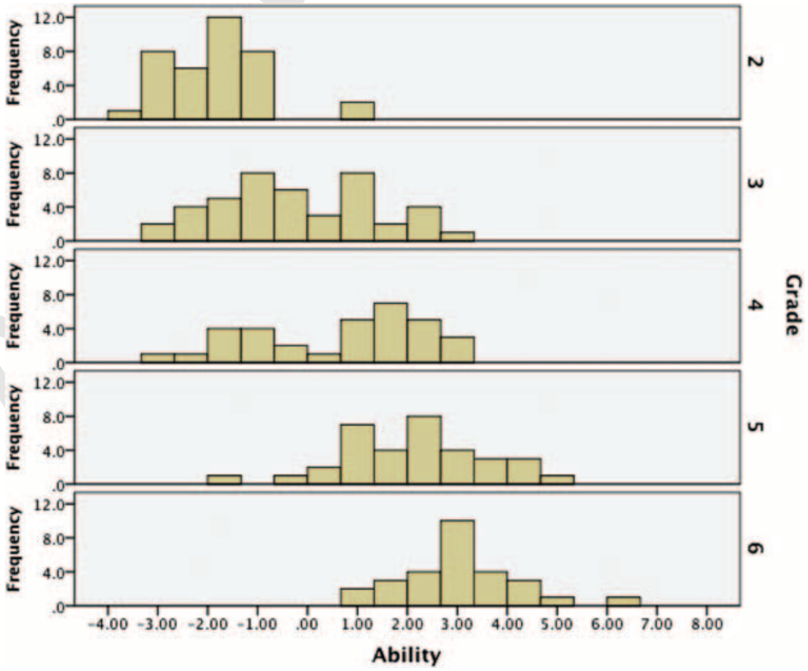


Figure 2. Distribution of ability estimates by grade level.

.79 and $r(83) = .80$, $ps < .01$, for Forms 1 and 2 respectively). This was true within each grade level as well. This positive correlation between our assessment and a general standardized math assessment provides some evidence of convergent validity.

Evidence based on response processes. Thus far, our analyses have focused on the accuracy of students' answers. However, past research and our construct map indicate that students' errors should not be random. Rather, an operational understanding of the equal sign as an indicator to "add up the numbers" should lead students either (a) to think the terms before the equal sign should add up to the term immediately after the equal sign (*add-to-equal*; e.g., answering 7 to 3 + 4 = □ + 5 or answering 4 to □ + 2 = 6 + 4) or (b) to think all the numbers in the equation should be added (*add-all*; e.g., answering 12 for either equation). To explore this, we coded children's errors on the six Level 2 and Level 3 equation-solving items based on their answers and their written work. Of the incorrect responses, 37% were nonresponses. Of the remaining errors, 63% were "add up the numbers" errors (52% add-to-equal errors and 11% add-all errors). The frequency with which children made "add up the numbers" errors was correlated with their estimated abilities on the assessment, $r(173) = -.57$, $p < .01$. Overall, students' errors often reflected an operational view of equivalence.

Characterizing Students' Knowledge Levels

Much of the power of IRT results from the fact that it models participants' responses at the item level. For example, 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 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 intense item-by-item error analysis.

Consider a student with the mean ability score of .71. This student would be expected to solve the Level 3 item 3 + 4 = □ + 5 (13SOL.L3.2) accurately 68% of the time and would be expected to solve few Level 4 items correctly. In contrast, a student with an ability score of −1.6 (1 *SD* below the mean) would be expected to solve this Level 3 item accurately only 17% of the time but would be expected to solve the easier Level 2 item 8 = 6 + □ (6SOL.L2.2) correctly 77% of the time. As we develop our measure over time, adding more and more items to the bank of known difficulty levels, this predictive power will grow in precision and generality.

Textbook Analysis

To help shed insight on the role of experience in the development of equivalence knowledge, we performed a textbook analysis of the textbook series used at the school, *Houghton Mifflin Math* (Greenes et al., 2005).

Method. Following the method used byMcNeil et al. (2006), we coded the equation structure surrounding each instance of the equal sign on every other page of the Grade 1–6 textbooks. Equation types are defined in Table 4. T4

Findings. Across grade levels, there was a steady increase in the number of instances of the equal sign per page (see bottom of Table 4). In first grade, the operations-equals-answer structure

Table 4
Textbook Analysis Results: Percentage of Instances of the Equal Sign in Each Equation Structure for Grades 1 Through 6

Structure type		Grade						
		1	2	3	4	5	6	Average
Operations-equals-answer structure		**97**	**82**	**70**	**52**	**38**	**31**	**62**
Unknown at end or no unknown	Operation(s) on left side of the equal sign and unknown quantity or answer on the right side (e.g., 5 + 2 = □)	91	75	48	35	18	11	46
Unknown on left side	Operation(s) and an unknown quantity on the left side (e.g., 4 + □ = 7)	6	7	22	17	20	20	15
Nonstandard equation structures		**0**	**10**	**24**	**41**	**59**	**68**	**34**
Operations on right side of equal sign	Operation(s) on right side of the equal sign and answer or an unknown quantity on the left side (e.g., 7 = 5 + 2)	0	4	3	2	18	11	6
No explicit operations	No explicit operation on either side of the equal sign (e.g., 12 in. = 1 ft., x = 4, 2/4 = 1/2, 3 = 3)	0	5	15	33	38	49	23
Operations on both sides of equal sign	Operations appear on both sides of the equal sign (e.g., 3 + 4 = 5 + 2)	0	1	6	6	3	8	4
No equation	Equal sign appears outside the context of an equation, such as in the directions (e.g., "Write <, >, or = to complete each statement")	3	7	6	6	3	1	4
Total instances of equal sign		492	363	696	671	859	1267	
Pages examined		320	310	314	314	311	309	
Instances per page		1.5	1.1	2.2	2.0	2.7	4.0	

dominated, accounting for 97% of all occurrences of the equal sign. There was a steady decrease in the frequency of this structure, with it eventually accounting for just 31% of occurrences in the sixth-grade text. In contrast, equations with no explicit operation (e.g., $1 \text{ foot} = 12 \text{ in}$) increased steadily from first to sixth grade, with this structure accounting for almost one half of the occurrences of the equal sign in sixth-grade text. The other structures were relatively rare, accounting for less than 15% of instances of the equal sign across the grades. Equations with operations on both sides were particularly rare, accounting for only 4% of instances overall.

We also inspected the textbooks for explicit definitions of the equal sign embedded either in a lesson or in the glossary. We found no explicit definitions. In the first-grade textbook, "equal sign" was included in the glossary, but the definition was simply an arrow pointing to the equal sign in the equation $2 + 3 = 5$. In the second-grade textbook, there was an entry for *equal to (=),* with the example $4 + 4 = 8$, 4 plus 4 is equal to 8. There was no entry in the third-grade or fifth-grade texts; the fourth-grade text did include an entry for *equal* in the glossary with the definition "having the same value," but there was no link to the equal sign. In the sixth-grade text, the definition of *equation* in the glossary was "A mathematical sentence that uses an equal sign to show that two expressions are equal. $3 + 1 = 4$ and $2x + 5 = 9$." This was the only definition that might support a relational definition of the equal sign.

Discussion

Numerous past studies have pointed to the difficulties elementary-school children have understanding mathematical equivalence (e.g., Behr et al., 1980; Falkner et al., 1999; McNeil, 2007; Perry, 1991; Rittle-Johnson & Alibali, 1999; Weaver, 1973), underscoring the need for systematic study of elementary-school students' developing knowledge of the topic. We used a construct-modeling approach to develop an assessment of mathematical equivalence knowledge. Our construct map specified a continuum of knowledge progression from a rigid operational view to a comparative relational view (see Table 1). We created an assessment targeted at measuring this latent construct and used performance data from an initial round of data collection to screen out weak items and to create two alternate forms of the assessment. The two forms of the revised assessment were shown to be reliable and valid along a number of dimensions, including good internal consistency, test–retest reliability, test content, and internal structure. In addition, our construct map was largely supported. Below, we discuss the strengths and weaknesses of our construct map, possible sources of increasing equivalence knowledge, benefits of a construct-modeling approach to measurement development, and future directions.

Construct Map for Equivalence

Describing children as having an operational or relational view of equivalence is overly simplistic. Rather, items of a broader range of difficulty can be used to capture students in transition between the two views (Level 2: Flexible operational) and to capture comparative reasoning based on equivalence ideas (Level 4: Comparative relational). As predicted by the construct map, children became increasingly flexible in dealing successfully with equation structures, and the structure of the equation had a large influence on performance. In contrast, the item class had limited influence on performance. For example, success evaluating versus solving a particular equation structure was often similar, and one class of items was not consistently easier than another. Our construct map for increasingly sophisticated abilities to deal with different equation structures across several item classes allows for a higher resolution description of children's knowledge of equivalence than was possible in previous studies.

Another benefit of a construct-modeling approach is that it encourages iterative refinement of the theoretical construct map in response to empirical findings. Indeed, the current findings suggest several potential refinements of the construct map. First, more attention should be paid to how the equal-sign definition items relate to performance on equation-structure and equation-solving items, as they were less likely to be at the expected level of difficulty than the other items. Of most note, generating a relational definition of the equal sign was much harder than solving or evaluating equations with operations on both sides. Rather, generating a relational definition was as hard as recognizing that a relational definition is the best definition of the equal sign (a Level 4 item). Past research has also found that explicit, verbalized knowledge of a relational definition of the equal sign takes longer to develop than the ability to solve or evaluate equations with operations on both sides (Denmark, Barco, & Voran, 1976; Kieran, 1981; Rittle-Johnson & Alibali, 1999). Likely, this definition item should be considered a Level 4 item. Further, Levels 3 and 4 may be more appropriately labeled as an *implicit relational view* versus an *explicit relational view*.

In addition, it may make sense to make finer grain distinction at Level 4. Compensation items were easier than items requiring more explicit thinking about the properties of equality linking the two sides of the equation. That is, children were more adept at employing the properties of equality (e.g., judging $89 + 44 = 87 + 46$ to be true without computing) than they were at explicitly recognizing or explaining those properties (e.g., recognizing and justifying that if $56 + 85 = 141$ is true, $56 + 85 - 7 = 141 - 7$ is also true). If the relative difficulty of these items persists in future studies, it may be worth distinguishing two sublevels to a comparative relational view.

Developing Knowledge of Equivalence

What causes children to develop increasingly sophisticated knowledge of equivalence? This study did not address this issue directly, as we did not manipulate children's experiences with equivalence ideas or directly observe classroom instruction. Teacher reports and a textbook analysis, however, provide some information that is informative. First, consider the potential role of exposure to different equation structures in textbooks. Textbooks heavily influence what children are exposed to in classrooms (Reys, Reys, & Chavez, 2004; Weiss, Banilower, McMahon, & Smith, 2001). Analysis of the participating students' textbooks indicated that exposure to nonstandard equation structures did increase dramatically with grade, accounting for 3% of instances of the equal sign in the first-grade textbook and 68% of instances in the sixth-grade text. A vast majority of these instances had no explicit operations (e.g., 1 foot = 12 in, 1/2 = 2/4), and the frequency of equations with operations on both sides was low across grades. Note that these nonstandard equation structures were much more prevalent in the sixth-grade textbook that we analyzed than in the four sixth-grade textbooks analyzed by McNeil et al. (2006; 30%–51% of instances). There appears to be large variability in presentation of nonstandard equation structures across textbook series.

Students' knowledge was developing earlier than would be predicted by mere exposure. For example, many children in Grade 2 were successful on items with operations on the right or no operations even though they were rarely exposed to these equation structures in their textbooks. Similarly, many older children were successful on items with operations on both sides, although these items were rare in their textbooks. A recent textbook analysis of a sixth-grade textbook from each of four countries (China, Korea, Turkey, and the United States) also suggests that simple exposure to nonstandard equation structures is not the primary sources of improving equivalence knowledge. The frequency of nonstandard equation structures was comparable in the textbooks from the four countries, even though students in China and Korea were much more likely to solve equations with operations on both sides correctly (Capraro, Yetkiner, Ozel, & Capraro, 2009).

It may be that, rather than simple textbook exposure, explicit attention to ideas of equivalence in classroom discussion, with attention to the equal sign as a relational symbol, is what promotes knowledge growth in this domain. Second-grade teachers in the current study reported discussing the meaning of the equal sign for about a week, in addition to exposing students to nonstandard equation structures. Such explicit attention to the meaning of the equal sign in second grade was not directly supported by the textbook but may reflect awareness by the second-grade teachers about the difficulty of this topic. These classroom discussions may have helped children gain a more flexible, albeit operational, view of equivalence. Teachers in fourth through sixth grade reported spending at least 3–5 days on solving equations with variables, and it is possible that attention to equation solving aided growth of a relational view of equivalence. We did not observe these classroom activities and discussions, but they are in line with teaching experiments on the effectiveness of classroom discussions of nonstandard equation structures and what the equal sign means (e.g., Baroody & Ginsburg, 1983; Jacobs et al., 2007).

It is also possible that cognitive differences, not just instructional ones, influence growth of equivalence knowledge across grades. For example, according to Case, older elementary-school children develop the ability to integrate two dimensions of a problem (Case & Okamoto, 1996), perhaps helping them coordinate the information on both sides of the equal sign. Children also become better able to inhibit task-irrelevant information through elementary school (e.g., Dempster, 1992), which perhaps helps them inhibit operational views of equivalence. These cognitive changes may, in part, explain improvements in equivalence knowledge with age. However, limitations in cognitive capacity do not prevent younger children from understanding equivalence when they are given extensive, well-structured instruction (Baroody & Ginsburg, 1983; Jacobs et al., 2007; Sáenz-Ludlow & Walgamuth, 1998).

Benefits of a Construct-Modeling Approach to Measurement Development

A construct-modeling approach to measurement development is a particularly powerful one for researchers interested in understanding knowledge progression, as opposed to ranking students according to performance. Although Mark Wilson has written an authoritative text on the topic (Wilson, 2005), there are only a handful of examples of using a construct-modeling approach in the empirical research literature (see Acton, Kunz, Wilson, & Hall, 2005; Masse, Heesch, Eason, & Wilson, 2006; Wilson, 2008), with only a few focused on academic knowledge (see Claesgens, Scalise, Wilson, & Stacy, 2009; Dawson-Tunik, Commons, Wilson, & Fischer, 2005; Wilson & Sloane, 2000). We found construct modeling to be very insightful and hope this article will inspire other educational and developmental psychologists to use the approach. This measurement development process incorporates four phases that occur iteratively: (a) proposal of a construct map based on the existing literature and a task analysis; (b) generation of potential test items that correspond to the construct map and systematic creation of an assessment designed to tap each knowledge level in the construct map; (c) creation of a scoring guide that links responses to items to the construct map; and (d) after administration of the assessment, use of the measurement model, in particular Rasch analysis and Wright maps, to evaluate and revise the construct map and assessment (Wilson, 2005). The assessment is then progressively refined by iteratively looping through these phases.

Another benefit of a construct-modeling approach is that it produces a criterion-referenced measure that is particularly appropriate for assessing the effects of an intervention on individuals (Wilson, 2005). We developed two versions of our equivalence assessment so that different versions could be used at different assessment points in future intervention or longitudinal research.

Our equivalence assessment could also help educators modify and differentiate their instruction to meet individual student needs. IRT can be used to assign ability scores, which teachers can use to classify children at different levels of equivalence knowledge. We found wide variability in performance within grades, and diagnostic information for individual students should help teachers differentiate their instruction to focus on items at the appropriate level of difficulty for a particular child. Differentiated instruction has been shown to improve student achievement (e.g., Mastropieri et al., 2006), but teachers often lack the tools for identifying students' knowledge levels and customizing their instruction (e.g., Houtveen & Van de Grift, 2001). Our measure of equivalence knowledge and the accompanying construct map could help facilitate this differentiation.

Future Directions and Conclusions

Although we have taken an important first step in validating a measure of equivalence knowledge, much still needs to be done. A critical next step is to provide evidence for the validity of the measure with a larger and more diverse sample. Such an effort will reveal whether items on the assessment function the same for different groups (e.g., grade levels or socioeconomic groups). It may be that our measure appears to be more cohesive than it actually is because we sampled students from a wide range of grades but were unable to test for the effect of grade on item functioning or dimensionality, given the limited number of students per grade level. We also need to know the predictive validity of the measure (e.g., does the measure help predict which students need additional math resources or who are ready for algebra in middle school?) Having a common assessment tool should also facilitate future efforts to better understand sources of changes in equivalence knowledge as well as to evaluate the effectiveness of different educational interventions.

References

Acton, G. S., Kunz, J. D., Wilson, M., & Hall, S. M. (2005). The construct of internalization: Conceptualization, measurement, and prediction of smoking treatment outcome. *Psychological Medicine, 35,* 395–408. doi:10.1017/S0033291704003083

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: Office of Vocational and Adult 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,* 221–246.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition, 44,* 75–106.

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 equals sign. *Mathematics Teaching, 92,* 13–15.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Capraro, R. M., Yetkiner, Z. E., Ozel, S., & Capraro, M. M. (2009, April). *An international perspective on sixth graders' interpretation of the equal sign.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school.* Portsmouth, NH: Heinemann.

Case, R., & Okamoto, Y. (1996). The role of central conceptual structures in the development of children's thought. *Monographs for the Society for Research in Child Development, 61*(12, Serial No. 246).

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93,* 56–85. doi:10.1002/sce.20292

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. W. (2005). The shape of development. *European Journal of Developmental Psychology, 2,* 163–195. doi:10.1080/17405620544000011

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review, 12,* 45–75. doi:10.1016/0273-2297(92)90003-K

Denmark, T., Barco, E., & Voran, J. (1976). *Final report: A teaching experiment on equality* (PMDC Report No. 6). Tallahassee: Florida State University.

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

Freiman, V., & Lee, L. (2004). Tracking primary students' understanding of the equality sign. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th conference of the International Group for the Psychology of Mathematics Education* (pp. 415–422). Bergen, Norway: Bergen University College.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

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

Greenes, C., Larson, M., Leiva, M., Shaw, J., Stiff, L., Vogeli, B., & Yeatts, K. (Eds.). (2005). *Houghton Mifflin math.* Boston, MA: Houghton Mifflin.

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

Houtveen, T., & Van de Grift, W. (2001). Inclusion and adaptive instruction in elementary education. *Journal of Education for Students Placed at Risk, 6,* 389–409. doi:10.1207/S15327671ESPR0604_5

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Jacobs, V. R., Franke, M. L., Carpenter, T., 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.

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

Kieran, C. (1992). The learning and teaching of school algebra. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390–419). New York, NY: Simon & Schuster.

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.

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

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. (2010). *A user's guide to Winsteps ministep Rasch-model computer programs.* Retrieved from http://www.winsteps.com/winman/index.htm?copyright.htm

MacGregor, M., & Stacey, K. (1997). Students' understanding of algebraic notation. *Educational Studies in Mathematics, 33,* 1–19. doi:10.1023/A:1002970913563

Masse, L. C., Heesch, K. C., Eason, K. E., & Wilson, M. (2006). Evaluating the properties of a stage-specific self-efficacy scale for physical activity using classical test theory, confirmatory factor analysis and item response modeling. *Health Education Research, 21*(Suppl. 1), 33–46. doi:10.1093/her/cyl106

Mastropieri, M. A., Scruggs, T. E., Norland, J. J., Berkeley, S., McDuffie, K., Tornquist, E. H., & Connors, N. (2006). Differentiated curriculum enhancement in inclusive middle school science: Effects on classroom and high-stakes tests. *The Journal of Special Education, 40,* 130–137. doi:10.1177/00224669060400030101

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

McNeil, N. M. (2007). *U*-shaped development in math: 7-year-olds out-perform 9-year-olds on equivalence problems. *Developmental Psychology, 43,* 687–695. doi:10.1037/0012-1649.43.3.687

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. doi:10.1111/j.1467-8624.2008.01203.x

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science, 28,* 451–466.

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,* 285–306. doi:10.1207/s15327647jcd0602_6

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,* 883–899. doi:10.1111/j.1467-8624.2005.00884.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

Mix, K. S. (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. C. (2006). Fostering relational thinking while negotiating the meaning of the equals sign. *Teaching Children Mathematics, 13,* 111–117.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Muthén, B., & Muthén, L. K. (2006). Mplus 4.2 [Computer software]. Los Angeles, CA: Author.

Muthén, L. K. (2004, November 2). Fit indices for categorical outcomes [Online forum comment]. Retrieved from http://www.statmodel.com/discussion/messages/23/26.html#POST3665

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

Oksuz, C. (2007). Children's understanding of equality and the equal symbol. *International Journal for Mathematics Teaching and Learning.* Retrieved from http://www.cimt.plymouth.ac.uk/journal/default.htm

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

Rasch, G. (1980). *Probabilistic model for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Reys, B. J., Reys, R. E., & Chavez, O. (2004). Why mathematics textbooks matter. *Educational Leadership, 61,* 61–66.

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

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Sáenz-Ludlow, A., & Walgamuth, C. (1998). Third graders' interpretations of equality and the equal symbol. *Educational Studies in Mathematics, 35,* 153–187. doi:10.1023/A:1003086304201

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* (pp. 161–187). Mahwah, NJ: Erlbaum.

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

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520. doi:10.1016/0010-0285(76)90016-5

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking.* New York, NY: Oxford University Press.

Steinberg, R. M., Sleeman, D. H., & Ktorza, D. (1990). Algebra students' knowledge of equivalence equations. *Journal for Research in Mathematics Education, 22,* 112–121. doi:10.2307/749588

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.

Warren, E. (2003, July). *Young children's understanding of equals: A longitudinal study.* Paper presented at the 27th conference of the International Group for the Psychology of Mathematics Education, Honolulu, HI.

Watchorn, R., Lai, M., & Bisanz, J. (2009, October). *Failure on equivalence problems is not universal.* Paper presented at the meeting of the Cognitive Development Society, San Antonio, TX.

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

Weiss, I. R., Banilower, E. R., McMahon, K. C., & Smith, P. S. (2001). *Report of the 2000 National Survey of Science and Mathematics Education.* Retrieved from the Horizon Research Inc. website: http://2000survey.horizonresearch.com/reports/status.php

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research, 8,* 1–22.

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

Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology, 216,* 74–88.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208. doi:10.1207/S15324818AME1302_4

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.

(*Appendices follow*)

Appendix A

Revised Assessment (Form 2) With Scoring Criteria for Select Items

Equation Structure Items (10 min)

1. "I'll show a problem for a few seconds. After I take the problem away, I want you to write the problem exactly as you saw it."

a. □ + 2 = 5

b. 5 + 2 = □ + 4

c. □ + 5 = 8 + 7 + 5

d. 74 + □ = 79 + 45 (dropped item)

Scoring criteria. Correct if all numerals, operators, equal sign, and unknown are in correct places. OK if numerals are incorrect.

2. For each example, decide if the number sentence is true. In other words, does it make sense? After each problem, circle True, False, or Don't Know.

a. 5 + 3 = 8 True False Don't Know (at ceiling)

b. 3 = 3 True False Don't Know

c. 31 + 16 = 16 + 31 True False Don't Know

d. 7 + 6 = 6 + 6 + 1 True False Don't Know

e. 5 + 5 = 5 + 6 True False Don't Know

3. For each example, decide if the number sentence is true. Then, explain how you know.

a. 7= 3 + 4 True False Don't Know

b. 6 + 4 = 5 + 5 True False Don't Know

Scoring criteria for explanations. Correct if mentions the word "same," that the inverse is true, or solves and shows both sides to be the same.

4. This problem has two sides. Circle the choice that correctly breaks the problem into its two sides.

8 + 2 + 3 = 4 + □

5. Without adding 67 + 86, can you tell if the statement below is true or false?

67 + 86 = 68 + 85. How do you know?

Scoring criteria for explanation. Correct if mentions relations between values on the two sides (e.g., "67 is one less then 68, same with 85 and 86").

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

56 + 85 = 141 is true.

Is 56 + 85 – 7 = 141 − 7 true or false?

How do you know?

Scoring criteria for explanation. Correct if mentions doing the same thing to both sides (e.g., "they subtracted 9 from both sides").

Equal Sign Items (5 min)

7. What does the equal sign (=) mean?

Can it mean anything else?

Scoring criteria. Correct if they give a relational definition, which mentions both sides being the same or equivalent. (Note: 30% of student who gave a relational definition did so only when prompted "Can it mean anything else?" They provided an operational or ambiguous definition on the first prompt.)

8. Which of these pairs of numbers is equal to 6 + 4? Circle your answer.

a. 5 + 5

b. 4 + 10

c. 1 + 2

d. none of the above

9. Which answer choice below would you put in the empty box to show that five pennies are the same amount of money as one nickel? Circle your answer.

a. 5¢

b. =

c. +

d. don't know

(*Appendices continue*)

10. Is this a good definition of the equal sign? Circle good or not good.

a. The equal sign means add. (at ceiling, not included) Good Not good

b. The equal sign means get the answer. (dropped item) Good Not good

c. The equal sign means the same as. Good Not good

11. Which of the definitions above is the *best* definition of the equal sign?

12. The equal sign (=) is more like: (dropped item)

a. 8 and 4

b. < and >

c. + and −

d. don't know

Equation-Solving Items (10 min)

DIRECTIONS: Find the number that goes in each box.

13. 3 + 4 = □ (at ceiling, not included)

14. 4 + □ = 8

15. 8 = 6 + □

16. 3 + 4 = □

17. □ + 2 = 6 + 4

18. 7 + 6 + 4 = 7 + □

19. 8 + □ = 8 + 6 + 4

20. 6 − 4 + 3 = □ + 3

DIRECTIONS: Find the number that goes in each box. You can try to find a shortcut so you don't have to do all the adding. Show your work and write your answer in the box.

21. 898 + 13 = 896 + □

22. 43 + □ = 48 + 76

23. Find the value of *n*. Explain your answer.

$n + n + n + 2 = 17$

Scoring criteria. For Items 13–23, answers within 1 of the correct answer were considered correct to allow for minor computation errors.

(*Appendices continue*)

Appendix B

Tetrachoric Correlation Matrix for Factor CFA

	4STR.L2	21STR.L3	22STR.L3	7STR.L2	20STR.L3	18STR.L3	3STR.L1	8STR.L2	9STR.L2	17STR.L3	19STR.L3	9STR.L4	27STR.L4	28STR.L4	30STR.L4
4STR.L1															
21STR.L3	0.453														
22STR.L3	0.267	0.504													
7STR.L2	0.274	0.372	0.588												
20STR.L3	0.508	0.416	0.606	0.691											
18STR.L3	0.068	0.221	0.344	0.509	0.79										
3STR.L1	0.243	0.37	0.515	0.717	0.68	0.638									
8STR.L2	0.511	0.489	0.397	0.634	0.839	0.62	0.659								
9STR.L2	0.473	0.459	0.236	0.513	0.723	0.544	0.578	0.976							
17STR.L3	0.459	0.401	0.555	0.749	0.887	0.859	0.869	0.797	0.715						
19STR.L3	0.404	0.381	0.56	0.747	0.913	0.818	0.868	0.839	0.853	0.994					
29STR.L4	0.417	0.41	0.324	0.413	0.582	0.622	0.498	0.498	0.667	0.433	0.633	0.574			
27STR.L4	0.516	0.258	0.45	0.529	0.795	0.564	0.811	0.637	0.553	0.749	0.728	0.566			
28STR.L4	0.387	0.512	0.579	0.54	0.716	0.578	0.669	0.641	0.734	0.807	0.838	0.539	0.921		
30STR.L4	0.417	0.301	0.369	0.272	0.514	0.622	0.314	0.582	0.294	0.453	0.315	0.476	0.566	0.583	
31STR.L4	0.188	0.293	0.501	0.283	0.517	0.598	0.491	0.459	0.457	0.651	0.604	0.531	0.546	0.639	0.928
10DEF.L3	0.178	0.335	0.385	0.721	0.718	0.53	0.708	0.598	0.546	0.771	0.766	0.694	0.636	0.596	0.413
1DEF.L1	0.157	0.366	0.452	0.591	0.607	0.429	0.659	0.715	0.732	0.757	0.749	0.456	0.784	0.641	0.178
5DEF.L2	0.192	0.363	0.4	0.639	0.421	0.391	0.525	0.45	0.311	0.548	0.586	0.371	0.476	0.623	0.371
11DEF.L3	0.136	0.165	0.407	0.662	0.66	0.544	0.658	0.612	0.562	0.617	0.706	0.509	0.422	0.312	0.362
23DEF.L4	0.089	0.064	0.332	0.383	0.529	0.519	0.34	0.269	0.412	0.392	0.428	0.449	0.37	0.37	0.361
2SOL.L1	0.34	0.123	0.229	0.477	0.177	0.095	0.312	0.344	0.465	0.376	0.333	-0.025	0.193	-0.055	-0.025
6SOL.L2	0.187	0.334	0.294	0.517	0.489	0.528	0.669	0.71	0.631	0.725	0.681	0.533	0.524	0.387	0.533
12SOL.L3	0.304	0.453	0.534	0.698	0.887	0.746	0.765	0.746	0.659	0.918	0.902	0.734	0.68	0.634	0.451
13SOL.L3	0.312	0.353	0.454	0.738	0.873	0.827	0.845	0.81	0.732	0.963	.939	0.666	0.73	0.696	0.496
15SOL.L3	0.275	0.418	0.458	0.592	0.883	0.775	0.86	0.736	0.656	0.911	0.887	0.643	0.799	0.798	0.436
14SOL.L3	0.419	0.526	0.42	0.693	0.907	0.778	0.887	0.825	0.675	0.935	0.875	0.653	0.707	0.749	0.521
16SOL.L3	0.109	0.414	0.461	0.565	0.79	0.701	0.737	0.592	0.479	0.826	0.787	0.545	0.676	0.761	0.545
25SOL.L4	0.527	0.338	0.574	0.678	0.876	0.736	0.862	0.792	0.613	0.954	0.871	0.548	0.721	0.662	0.589
24SOL.L4	0.577	0.378	0.439	0.702	0.826	0.703	0.823	0.742	0.617	0.927	0.865	0.607	0.782	0.74	0.445
26SOL.L4	0.476	0.719	0.333	0.418	0.55	0.447	0.452	0.665	0.537	0.622	0.598	0.467	0.705	0.605	0.424

(*Appendices continue*)

Appendix B (*continued*)

	31STR.L3	10DEF.L3	1DEF.L1	5DEF.L2	11DEF.L3	23DEF.L4	2SOl.L1	6SOL.L2	12SOL.L3	13SOL.L3	15SOL.L3	14SOL.L3	16SOL.L3	25SOL.L4	24SOL.L4
10DEF.L3	0.423														
1DEF.L1	0.334	0.598													
5DEF.L2	0.526	0.422	0.594												
11DEF.L3	0.457	0.798	0.373	0.493											
23DEF.L4	0.641	0.496	0.212	0.128	0.813										
2SOL.L1	-0.244	0.16	0.562	0.275	-0.024	-0.066									
6SOL.L2	0.311	0.486	0.52	0.488	0.427	0.146	0.483								
12SOL.L3	0.593	0.751	0.673	0.486	0.692	0.437	0.1	0.612							
13SOL.L3	0.675	0.835	0.767	0.496	0.732	0.459	0.349	0.698	0.95						
15SOL.L3	0.698	0.752	0.783	0.487	0.586	0.385	0.323	0.587	0.851	0.897					
14SOL.L3	0.578	0.795	0.867	0.562	0.713	0.322	0.245	0.694	0.921	0.978	0.938				
16SOL.L3	0.728	0.698	0.739	0.461	0.7	0.344	0.287	0.631	0.834	0.885	0.828	0.929			
25SOL.L4	0.792	0.668	0.792	0.542	0.613	0.407	0.203	0.537	0.883	0.951	0.837	0.862	0.798		
24SOL.L4	0.446	0.636	0.8	0.61	0.667	0.34	0.144	0.597	0.855	0.919	0.818	0.816	0.808	0.908	
26SOL.L4	0.511	0.548	0.752	0.625	0.212	0.271	0.155	0.48	0.542	0.531	0.619	0.652	0.498	0.626	0.558

Received June 30, 2009
Revision received August 13, 2010
Accepted August 25, 2010 ■

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES **1**