University of Alberta

Improving Children's Understanding of Mathematical Equivalence

by

Rebecca Patricia Dianne Watchorn

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology

Spring 2011
Edmonton, Alberta

Dedication

To Emma, Ali, and the rest of the next generation of students, with the hope that we can help make mathematics make sense to all students. To my dad for instilling a love of math in me, and to Gram, who loves learning but never loved math.

Abstract

A great majority of children in Canada and the United States from Grades 2-6 fail to solve equivalence problems (e.g., 2 + 4 + 5 = 3 + __) despite having the requisite addition and subtraction skills. The goal of the present study was to determine the relative influence of two variables, instructional focus (procedural or conceptual) and use of manipulatives (with or without), in helping children learn to solve equivalence problems and develop an appropriate understanding of the equal sign. Instruction was provided in four conditions consisting of the combination of these two variables.

Students in Grade 2 (n = 122) and Grade 4 (n = 151) participated in four sessions designed to assess the effectiveness of four instructional methods for learning and retention. Session 1 included a pretest of equivalence problem solving and three indicators of understanding of the equal sign. In Sessions 2 and 3 instruction was provided in one of the four instructional conditions or a control condition. Students were tested for their skill at solving equivalence problems immediately following instruction and at the beginning of Session 3 to assess what they had retained from Session 2. In Session 4, one month later, children were re-tested on all of the tasks presented in Session 1 to assess whether instruction had a lasting effect.

All four instructional groups outperformed the control group in solving equivalence problems, but differences among instructional groups were minimal. Performance on indicators of understanding, however, favoured students who received conceptually focused instruction. Preliminary evidence was found that

children's understanding of problem structure and attentional skill may be associated with the ability to benefit from instruction on equivalence problems. Children clustered into four groups based on their performance across tasks that are consistent with the view that children's understanding of the equal sign develops gradually, beginning with learning the definition.

These findings suggest that a relatively simple intervention can markedly improve student performance in the area of mathematical equivalence, and that these improvements can be maintained over a period of time and show some limited generality to other indicators that children understand equivalence.

Acknowledgement

First and foremost, my sincere gratitude goes to my supervisor, Dr. Jeffrey Bisanz, for his exceptional mentorship throughout this project and my entire graduate career. Jeff cared about every aspect of my projects and was always open for consultation, whether it was during "Math Group," over a pizza dinner with his family, or a late-night panicked phone call. Jeff, I truly appreciate your wisdom, dedication, and enthusiasm that you have shared so generously. As with all of his students, Jeff cared about me as a whole person, and cultivated not only my academic but citizenship skills as well. You pushed me to reach beyond the limits I put on myself and for this I will be forever grateful.

I would also like to thank Drs. Connie Varnhagen and Patricia Boechler for serving on my supervisory committee. Connie and Patricia offered support and guidance at every step of my journey, and opportunities for growth in their labs. I was fortunate to be surrounded by such knowledgeable and approachable mentors.

The members of the ever-evolving Mathematical Cognition Lab provided immeasurable support and new life-long friends. Thanks to Drs. Carmen Rasmussen and Jody Sherman LeVos for welcoming me to my new home, and for helping me to navigate through the maze of a new program and department. A special thanks to Jody for introducing me to equivalence problems! Carley Piatt has been the best officemate and personal/career counselor a PhD student could ask for and the multi-talented Anna Matejko has been an inspiration with her passion for learning.

Along with Jeff and Jody, Dr. Helena Osana made significant contributions to the design of this project and Maysa Taha and Jessica Sachs-Cardinal assisted with the implementation. Not only were Maysa and Jessi incredibly efficient, professional, and dedicated, but they were also extraordinarily fun and friendly people, making each day I worked with them a pleasure.

This project would not have been possible without the support of the Black Gold Regional Schools, particularly the teachers who let us disrupt their class time and time again. And of course, thanks to the many students who worked so diligently on each of the "crazy" math problems we gave them!

Funding support was provided in various forms from the Natural Sciences and Engineering Research Council (NSERC), the Social Sciences and Humanities Research Council (SSHRC), and the Alberta Ingenuity Fund.

I have to thank not only those directly impacted my research, but also those who supported me in a broader sense. Thank you to: Christine Boyle in the Psychology General Office for greeting me each day with a smile and making me feel valued; Isaac Lank and Philip May for extraordinary tech support; Dr. Elena Nicoladis for serving on my examining committees and supporting my teaching endeavours; Drs. Nancy Galambos and Roger Dixon for hosting wonderful welcoming parties and creating a sense of community in the Developmental Science area; Dr. Tom Johnson for keeping me company on “the evening shift”; all of the incredible friends I have made through the U of A Ringette Team and the U of A Triathlon Club for keeping me fit, sane, and laughing through everything; Sabrina, Carolyn, Erika, DeeAnn, Sandy, and Katie for emotional and spiritual support any time of the day or night, "I've learned that true friendship continues to grow, even over the longest distance"; my extended family at CVUC for your ever-encouraging words and always reminding me that people back home were thinking of me; Aunt Kathy and Uncle Geoff for always welcoming me “home” with open arms and an open refrigerator; my brother and my grandparents for their complete faith in me; my parents for their continuous support, including Mom agreeing to write me daily updates to get me out of bed in the morning!; and finally to Pat, for your unparalleled patience and loving me unconditionally.

Thank you, all, for making this an incredible journey.

Table of Contents

List of Tables

List of Figures

Improving Children's Understanding of Mathematical Equivalence

Mathematics is a skill people use nearly every day of their lives, often without even being aware of it. Math is used in determining which purchase is a better deal, working out travel distances or times, thinking about speed, understanding sports statistics, interpreting directions for assembling furniture, and reading a graph about the national budget, to name only a few examples (Bryant & Nuñes, 2002; DeLoache, 2002). Math is a specialized language used to communicate particular kinds of ideas (Carpenter, Franke, & Levi, 2003). A person needs a certain amount of fluency in this language to be able to function effectively in modern society and to have access to a large portion of the job market. Because technologies and job requirements are rapidly changing, we cannot anticipate the specific skills that today's children will need in their careers. Instead, we need to prepare students to be able to adapt their skills and acquire new knowledge as they encounter new problems (Ginsburg, Lee, & Boyd, 2008). The language of mathematics is more than simply a collection of ways to get answers (Carpenter & Lehrer, 1999), and the symbolism of math is the hardest form of language for children to learn (Ginsburg et al., 2008). Mathematics education is also recognized as having long-term economic benefits because of its importance for the future work force (Ginsburg et al., 2008). Because math is used so frequently in people's personal and professional lives, the development of mathematical thinking is an important domain of study. The development of mathematical thinking can also be used as a vehicle to study cognitive development.

Competence in many areas of math is important, but algebra has been highlighted as an especially important area. Many refer to algebra as the "gatekeeper" to higher math (Carpenter et al., 2003) and Moses and Cobb (as cited in Jacobs, Franke, Carpenter, Levi, & Battey, 2007) compare people today without algebra to people who could not read and write in the industrial age. Unfortunately, difficulty with algebra is extremely common (Blume & Heckman, 1997; Schmidt, McKnight, Cogan, Jakwerth, & Houang, 1999) and most students never acquire any sense of the structural aspects of algebra (Kieran, 1992).

Both teachers and researchers have known for decades that one of the most significant hurdles students face in learning math is the transition from arithmetic to algebra (Carpenter, Levi, & Farnsworth, 2000). For many people, adults included, arithmetic and algebra are separate collections of meaningless procedures and manipulations of numbers and symbols. In fact, both are based on specific principles. It is through understanding these principles that one develops an appropriate mathematical foundation in which both arithmetic and algebra make sense (Bisanz, Sherman, Rasmussen, & Ho, 2005; Carpenter et al., 2003).

A significant difficulty that many students face in relating arithmetic and algebra emerges from a lack of understanding of the relations expressed by number sentences (Falkner, Levi, & Carpenter, 1999). For example, "4 + 5 = 4 + 4 + 1" represents a mathematical relation that is crucial to both arithmetic and algebra. The concept of equality is an important component in developing algebraic reasoning in children (Falkner et al., 1999). However, many children develop an *operator* view of the equal sign, interpreting the symbol as meaning, "put the total next" or "add up all the numbers", instead of a *relational* view expressing the equality of both sides of an equation. Students with an operator view of the equal sign generally hold two additional misconceptions: (a) that all operations must be performed on all of the given numbers and (b) that the operations must be on the left of the equal sign and the answer on the right. Together, these three misconceptions are called an "addition schema" (McNeil & Alibali, 2002) or "operational patterns" (McNeil & Alibali, 2005b). Students with these misconceptions have difficulty transitioning to algebra where problems do not necessarily take this form (Kieran, 1981; Knuth, Stephens, McNeil, & Alibali, 2006). The added cognitive strain children experience when they hold these misconceptions may make it difficult for them to focus on other new concepts, such as variables, that are introduced when learning algebra.

The equal sign is a very important component of high school algebra (Kieran, 1981), but it appears that some students continue to misinterpret the meaning of the sign well into high school and even beyond (Baroody & Ginsburg, 1983; Byers & Herscovics, 1977). High school algebra requires students to be continuously transforming equations into equivalent statements. To perform this task, one must understand that if the same operations are performed on both sides of the equal sign, the relation between the two expressions is unaltered. If a person does not have a relational interpretation of the equal sign, these transformations make little sense and can only be memorized as rules.

The difficulty of the transition to algebra is well recognized by researchers and many school districts, who see the need to work with children from a young age and over an extended period of time to develop a solid foundation for algebra (Carpenter et al., 2003; Falkner et al., 1999; Jacobs et al., 2007). Focusing on the fundamental properties builds a strong foundation for students to be able to think in flexible and powerful ways in the future. If we could determine the factors involved in helping children develop an appropriate understanding of the equal sign, we could use that information to help design instructional materials for teachers, allowing the teachers to best prepare their students for algebra and higher math.

Performance on equivalence problems encompasses many aspects of cognitive development such as perception, attention, memory, learning, and concepts. Through studying how children solve equivalence problems, we have the potential to learn about each of these aspects, how they are interrelated, and how they are integrated in the larger context of a child's development. Insights into the development, coordination, and organization of human cognition can be gleaned from advancing our understanding of the knowledge and skills that support arithmetic performance (Bisanz et al., 2005).

Experiments manipulating performance on equivalence problems have generated numerous insights into the specific abilities and limitations children possess and how these systems develop. I begin this paper by reviewing the empirical findings related to equivalence, and then I present a study designed to identify factors that contribute to improving children's understanding of mathematical equivalence.

Review of Empirical Findings

Success rates. The first step in developing a model of any phenomenon is to accurately delineate the performance observed, both the successes and failures, so that we have a clear picture of the performance and so that boundary conditions can be determined. This process is important for making inferences about the mechanisms involved in development (Bisanz, Morrison, & Dunn, 1995).

The low success rate of elementary school children in North America on equivalence problems has been well documented. McNeil and Alibali (2005b) found that without instruction about 74% of children aged 7 to 11 years of age failed equivalence problems. Falkner et al. (1999) reported that all 145 Grade 6 students in a single school answered equivalence problems with either an "add-all" strategy (adding all of the numbers presented in the problem and putting the total on the blank) or an "add-to-equal" strategy (adding up the numbers to the left of the equal sign and putting the total on the blank). Even in middle school children have difficulty with the equal sign. Oksuz (2007) found only 8% of the fifth grade and 25% of the sixth grade students at an urban U.S. elementary school provided a relational explanation of the equal sign. Even more surprisingly, 100% of the students said 3 + 8 = 11 X 6 = 66 – 5 = 61 was true.

Consistent with the poor performance across ages reported above, Falkner et al. (1999) and Behr, Erlwanger, and Nichol (1976) all found that performance on equivalence problems did not improve with age from Grades 1 through 6, nor did children of this age change their thinking about equality (Kieran, 1981). Capraro, Capraro, Ding, and Li (2007) found a similarly low success rate with

American Grade 6 students averaging about 28% correct. Performance among students of this age, however, is not necessarily consistent across different methods of presenting problems. Carpenter et al. (2003) found that some students who responded correctly to true/false number sentence questions (e.g., Is 8 + 2 = 5 + 5 true or false?) subsequently failed equivalence problems on which they were required to produce the answer.

Context-dependent performance is not unusual with equivalence problems. In fact, performance varies widely depending on whether problems are presented in a non-symbolic form (e.g., blocks) or symbolic form (e.g., Arabic numerals) (Sherman & Bisanz, 2009). Children's definitions of the equal sign also vary depending on whether the symbol is presented with coins, alone, or with numbers in an addition context (e.g., 4 + 3 = 7; Seo & Ginsburg, 2003).

Vergnaud, Benhadj, and Dussouet (as cited in Kieran, 1981) observed errors among 13-year-olds that are consistent with the children encoding procedures and not equalities. For example, the students would often record steps of a word problem by simply recording their procedures such as 1063 + 217 = 1280 – 425 = 1063. The problem here lies in the symbolic representation of their ideas. Even some high school students struggle with the meaning of the equal sign. Byers and Herscovics (1977) found that high school students did not consistently use the equal sign as a symbol of equivalence, and their errors and notation suggested that many still interpreted the equal sign as an operator (Kieran, 1981).

International performance. One explanation that has been proposed for why children perform so poorly is that perhaps children are predisposed to think of equality in terms of calculating answers rather than relating among quantities. This possibility seems unlikely because children as young as six years of age can learn the correct use of the equal sign, and their success is not dependent on computational skill (Carpenter et al., 2003). Performance in Asian countries has also been notably higher than in North America. Li, Ding, Capraro, and Capraro (2008) found that 98% of a Chinese sample of Grade 6 students solved equivalence problems correctly, versus 29% of a comparable US sample. In a similar study, Capraro et al. (2007) found Grade 6 Chinese students averaged 98% on four written problems. Although four problems is a small set on which to base conclusions, previous studies have shown success rates are similar whether children solve three (e.g., McNeil & Alibali, 2000), four (e.g., Rittle-Johnson & Alibali, 1999) or more than four math equivalence problems (McNeil, 2008). Watchorn, Lai, and Bisanz (2009) found that Taiwanese children who had recently completed Grades 2-4 averaged 84-88% correct on equivalence problems, and even children who had only recently completed Grade 1 averaged 51% correct. These results suggest that the great difficulty North American children experience with these problems is not universal and success is possible even at a young age.

Operator interpretation. Children's misinterpretation of the equal sign as an operator instead of a relational symbol is not new. It has been studied since at least 1976 (Behr et al., 1976) and acknowledgement of the problem was documented as early as 1932 (Renwick, 1932), yet the problem persists. Children typically define the equal sign as meaning “the answer comes next”, “use all the numbers”, or “add up all the numbers” (Carpenter et al., 2003). These interpretations reveal that they are thinking of the equal sign as a symbol that directs them to “do something”, similar to other operators such as the plus sign (+) or the minus sign (-). A more accurate view of the equal sign would be as a relational symbol, similar to the greater than (>) or less than (<) signs, indicating an equivalent relation between two expressions. The equal sign is also sometimes incorrectly used to extend a problem, where it simply indicates the result of the operations directly preceding it (e.g., 8 + 4 = 12 + 5 = 17). In these examples, the equal sign simply separates a problem and its answer (Kieran, 1981).

Even a relational view can lead to different strategies to determine whether two expressions are equivalent or to find the solution to a problem. One strategy is to compute the operations in each expression and compare the results, and the other is to examine the relations between the expressions. For example, with a problem such as 8 + 4 = __ + 5, one approach would be to add up both sides to get 12, and another would to be to reason that 5 is one more than 4, so the blank has to be one less than 8. Examining the relations between the expressions is a more sophisticated approach and lends the most value when transitioning to algebra, but it requires a thorough understanding of the equal sign.

Not only does a relational view ease the transition to algebra, but an operator interpretation of the equal sign has negative consequences. For example, students with an operator interpretation find it difficult to read arithmetic sentences that do not reflect the order of their calculations (Kieran, 1981). This makes it hard for these students to solve problems which would otherwise be easily within their reach.

As students learn more about algebra, they tend to progress from an operational to a relational view of the equal sign (Alibali, Knuth, Hattikudur, McNeil, & Stephens, 2007), but the progression does not occur as quickly as teachers predict (Asquith, Stephens, Knuth, & Alibali, 2007). McNeil and Alibali (2005a) found that students in Grades 3-5 held an operator view of the equal sign in all three contexts explored: equal sign alone, in a typical addition problem, and in an equivalence problem. But by Grade 7 students interpreted the sign as a relational symbol when presented in an equivalence problem. The same Grade 7 students still interpreted the equal sign as an operator when the sign was presented alone or in the context of an addition equation. This finding suggests that Grade 7 students may be in the process of transitioning from an operational to a relational view of the equal sign. Similarly, Baroody and Ginsburg (1983) suggested the period of transition happens at about 13 years of age, and Herscovics and Kieran (1980) found that 12- to 14-year-olds initially held an operator view of the equal sign but changed to a relational view after training.

As was mentioned above, children with an operator interpretation of the equal sign tend to adhere to three operational patterns: (a) the equal sign means the total, (b) all operations must be performed on all of the given numbers, and (c) all operations must be on the left side of the equal sign with the answer on the right. Among 7- to 11-year olds, adherence to operational patterns predicted change in strategy use following a simple (1-minute) lesson (McNeil & Alibali, 2005b). The lesson either involved asking the child to notice where the equal sign was in the problem, and/or explaining to the child that the equal sign means things on one side are the same amount as the things on the other side. Children who were less entrenched in the operational patterns generated more new strategies, more correct strategies, and achieved more success on transfer problems (McNeil & Alibali, 2005b).

Although undergraduate and graduate students appear to have a relational interpretation of the equal sign regardless of the context under which they are asked for a definition, their performance on equivalence problems under some circumstances indicates they may still occasionally view the equal sign in an operational way. McNeil and Alibali (2005b) found that when operational patterns had been activated and the participants were presented equivalence problems only very briefly (1.5 s), undergraduates were more likely to use the add-all strategy to solve the equations than when they are given an unlimited amount time. This suggests that under pressure even well-educated adults may revert to an operator interpretation.

Further to this finding, McNeil, Rittle-Johnson, Hattikudur and Petersen (2010) found that, even when not under time pressure, undergraduates who had first solved addition facts were less likely to correctly solve equivalence problems than students who first participated in control activities. It appears that an operational view of the equal sign is frequently not entirely abandoned even among undergraduates.

Problem perception. The misinterpretation of the equal sign may only be one indication of a larger problem. It appears that children interpret an expression such as 2 + 4 as an instruction to do something, even if the equal sign is not present (Behr, Erlwanger, & Nichols, 1980). They perceive problems to be in the form __ + __ = __ even when they are not, and will rearrange numbers or cross out signs to order the problem "correctly." One child illustratively expressed the problem by explaining to the experimenter, " __ = 2 + 5" is 'backwards'" and went on to ask, "Do you read backwards?" This example demonstrates the child's expectation for equations to appear in a canonical form. Similar to the operational patterns described earlier (McNeil & Alibali, 2005b), Baroody and Ginsburg (1983) noted that children expect written (horizontal) equations to take a particular form: two or more terms on the left, the result on the right, and the equal sign in between as a connecting symbol (Baroody & Ginsburg, 1983). When a person only uses one strategy to solve a problem, the person only needs to encode the problem features relevant to that strategy (McNeil & Alibali, 2004), which can optimize performance. However, failing to encode other problem features can prove problematic, especially if the person encounters problems that cannot be solved with the strategy he has been using. If children only perceive the problem features relevant to the typical addition schema on which they have been trained, they may be at great risk for difficulties in later years (McNeil & Alibali, 2002). Subsequent learning on equivalence problems is positively related to knowledge of which features of problems to encode (McNeil & Alibali, 2005b).

Success and failure of interventions. If children's difficulty with equivalence problems were simply due to lack of attention to the details of the problem, asking children to note where the equal sign is in the equation and asking them to point to it should improve performance. This technique, however, does little to improve performance (McNeil & Alibali, 2002; 2005b). If children misinterpret the meaning of the equal sign, simply drawing attention to the sign will not remedy the problem. One might presume that to correct the misconception, all that would be needed would be to explain to the children the correct meaning of the equal sign. Unfortunately, it is not so straightforward. Children do not easily relinquish the beliefs they have developed over years, and explaining the correct meaning is not enough to convince them to change their thinking (Carpenter et al., 2003). Even in kindergarten when the children have presumably had the least exposure to the sign, one or two examples or a simple explanation is not enough to eliminate the misconception about the equal sign, (Falkner et al., 1999). Explicitly telling children that the equal sign means that whatever is on one side has to be the same amount as whatever is on the other side was moderately helpful with children in Grades 4 and 5 (Perry, 1991; Watchorn & Bisanz, 2005), but not as helpful with younger children in Grades 2 and 3 (McNeil, 2007; Watchorn & Bisanz, 2005) or Grades 3 and 4 (Alibali, 1999; McNeil & Alibali, 2005b). McNeil (2008) found that even after several lessons on mathematical equivalence from the classroom teacher, children solved an average of only 0.73 out of 12 ($SD = 0.98$) problems correctly.

Falkner et al. (1999) reported a case in which a Grade 2 teacher had a class discussion and told the class she agreed with two students who gave correct explanations of the equal sign, but still many students did not change their thinking about the meaning of the sign. Seo and Ginsburg (2003) also suggested that exposing children to the equal sign in contexts other than in typical addition problems might not be enough, but it is possible that the alternate contexts to which they exposed children were not of the appropriate kind for children to develop a relational view in arithmetic. They presented the equal sign with coins or with rods, but they did not try to combine alternate presentations with Arabic numbers.

Perry (1991) examined the effect of principle versus procedural instruction. She found that instructing students about the principle helped students' performance on a variety of problems, but instruction on a specific procedure decreased their chance of success on problems that differed from the type on which they were instructed. Performance was hindered on the transfer problems when procedural instruction was given, even if it was given in combination with principle instruction. The children tended to focus on the steps of the procedure and not why the procedure worked.

Clearly, verbally explaining the meaning of the sign to children is not enough. Children need to be able to link this new knowledge to existing knowledge. Many researchers have tried to use analogical reasoning to help children connect the ideas they are learning about equivalence to ideas they already have. Alibali (1999) used the verbal analogy of a teeter-totter with Grade

3 and 4 students, and the mean proportion of correct answers on a paper-and-pencil posttest improved from 0% on the pretest to 25% on the posttest. This improvement was a significantly larger improvement than a control group that only received feedback, but 25% correct is still very low. Watchorn and Bisanz (2005) found similar results with Grade 2 students, but Grade 4 students performed significantly better with an average of 60% correct following instruction. This improvement was maintained over a two-week interval. In a separate condition Watchorn and Bisanz provided instruction with a balance scale and found that for Grade 2 students, this method was significantly better than the verbal analogy or an explanation of what the equal sign is and what the goal is of equivalence problems. The balance scale approach appears to be somewhat more effective than some other methods with younger children, as Denmark, Barco, and Voran (1976) also found it to be effective even with Grade 1 students. Children who were exposed to different equation forms via the balance scale activities (e.g., 6 = 4 + 2) were more likely to accept non-canonical presentations, but the children still primarily interpreted the equal sign as an operator.

These discouraging results have led some to conclude that intellectual development or general conceptual limitations in childhood are contributing factors to children viewing the equal sign as an operator (Denmark et al., 1976). But, as mentioned above, there are some cultures where this problem is relatively minor in comparison to North America, and thus it is unlikely that cognitive limitations prevent children of this age from achieving success on these problems. Fortunately, some interventions have achieved slightly better results. For example, following a simple lesson, encouraging Grade 3-5 children to explain the reasoning behind their solution improved mean performance from approximately 5% to 79% correct, but did not affect other measures of conceptual knowledge about mathematical equivalence (Rittle-Johnson, 2006). Giving children goals to strive for improved the performance of approximately 75% of students on conceptual knowledge measures, but did not affect performance on equivalence problem solving (McNeil & Alibali, 2000). Having teachers include informative gestures in their instruction led to approximately 47% of students adopting the strategy demonstrated by the teacher (Goldin-Meadow, Kim, & Singer, 1999). Incredibly, forcing children to produce gestures corresponding to a correct solution also significantly improved their performance, however the mean score on problem solving was still only 40% (Goldin-Meadow, Wagner Cook, & Mitchell, 2009).

Researchers exploring performance on mathematical word problems have found that children learned more efficiently and deeply when worked examples were interleaved with practice problems, rather than simply solving the same problems on their own (see Atkinson, Derry, Renkl, & Wortham, 2000, for a review). This approach holds promise in equivalence instruction. Similarly, Rittle-Johnson and Star (2007) found that generating conceptual explanations for partners improved Grade 7 students' own learning of how to solve algebraic equations. Matthews and Rittle-Johnson (2009), however, found no effect for self-explanation prompts when instructing Grade 3-5 students about how to solve equivalence problems.

Some of the most successful techniques are centered on intensive professional development. Carpenter et al. (2000) had relatively good success with a study that built on Carpenter and Levi's Cognitively Guided Instruction (CGI) and focused on teacher professional development throughout the school year. Teachers participated in summer workshops and meetings throughout the year to learn how to analyze the structure and basic properties of arithmetic, consider learning contexts, and consider how students might think about specific problems. Two-hundred forty elementary students participated in the study, which included several class discussions about true/false number sentences that challenged misconceptions about the equal sign. By the end of the year, Grade 1 and 2 students averaged 66% correct on equivalence problems, Grade 3 and 4 students averaged 72% correct, and Grade 6 students averaged 84% correct. Falkner et al. (1999) report another success story where a Grade 2 class participated in many discussions throughout the year, and by the end of the school year most children held a relational view of the sign. Sáenz-Ludlow and Walgamuth (1998) reported similar success with a group of 14 Grade 3 students considered to be "at-risk." Jacobs et al. (2007) conducted a professional development project with 180 teachers in one of the lowest performing school districts in California. The content of their program focused on relational thinking. The teachers were encouraged to help the children understand the equal sign as an indicator of a relation, to use number relations to simplify calculations, and to verbalize conjectures about why solutions were correct or incorrect.

Students in the participating teachers' classrooms scored significantly better than students from other classrooms, but the highest group mean was still only 60%.

Baroody and Ginsburg (1983) investigated the effect of the Wynroth Curriculum (Wynroth, 1975), which makes a concerted effort to encourage a relational rather than an operator view of "equals." This individualized approach consists of a sequence of games in which learning the rules of the game teaches the child mathematical concepts. The approach was quite successful, as most of the 45 Grade 1-3 students in the study accepted atypical equations as correct if they were in a form to which they had been exposed in the games (e.g., 13 = 7 + 6), and about 50% of the students were able to correctly judge the sensibility of equation forms to which they had not been exposed (e.g., 7 + 6 = 14 – 1 or 2 + 3 = II + III). One drawback of these approaches is that they depend on intensive professional development, which is often not practical for widespread adoption, at least in the short-term. Widespread adoption would require a substantial investment of time and money in addition to significant changes to the current pedagogical approaches.

Fortunately, another approach is also showing promise. In Asia, the equal sign is introduced in the context of relations and teachers are more likely to highlight its relational meaning (Capraro et al., 2007). Asian textbooks are also more likely to present the equal sign in a manner that helps children develop a relational view of the sign (Li et al., 2008). In North America, comparing the equal sign to relational symbols and exposing children to atypical equations has been moderately successful as an intervention method for children who have developed the operational view of the sign. Hattikudur and Alibali (2007) found that Grade 3 and 4 students who compared the equal sign to inequality symbols performed better on a posttest of conceptual understanding than students who learned only about the equal sign. Their performance on equivalence problems, however, did not improve significantly. McNeil (2008) found that Grade 2 and 3 students were more successful on equivalence problems following lessons about the equal sign that used atypical equations (e.g., 28 = 28) as examples than they were following lessons that used typical equations (e.g., 15 + 13 = 28), however performance was still extremely low in both groups (means scores of less than 8% correct).

Although they only had a very small sample of 6 participants, Herscovics and Kieran (as cited in Kieran, 1981) were able to extend most of their 12- to 14-year-old students' use of the equal sign to include equivalence equations by providing lessons with equations that included multiple operations on both sides of the equal sign. One student, however, insisted on writing 4 + 3 = 6 + 1 as 4 + 3 = 7 and another insisted on inserting the "answer" between both sides (5 X 3 = 15 = 10 + 5). Thus, although promising, this technique was not completely successful.

Anderson (cited in Baroody & Ginsburg, 1983) also reported that explicitly teaching Grade 2 children to treat the equal sign as a relational symbol increased their likelihood to accept atypical equation forms. Carpenter et al. (2003) found that it was useful to use words that express the relation more directly ("is the same amount as"), and to use notation that shows that the numbers on the two sides of the equal sign represent the same numerical value. They concluded that using the equal sign in ways that do not represent a relation between numbers may mislead children, and that repeated exposure to a variety of correct forms may help children solidify their conception of what the equal sign means.

Varied performance depending on problem type. When discussing children's understanding of equivalence, an important consideration is how we define "understanding." Sometimes individuals appear to have knowledge of a concept in one context but not in another (Bisanz & LeFevre, 1992; Bisanz, Watchorn, Piatt, & Sherman, 2009; McNeil & Alibali, 2005a). Frequently the demonstration of knowledge is dependent on how we elicit the knowledge. For example, a learner might be able to correctly evaluate someone else's solution but have difficulty producing that solution on her own (Bisanz & LeFevre, 1992). Schneider and Stern (2010) also highlighted the difficulty in obtaining an accurate picture of ability through only one task. For these reasons it is important to examine the different contexts within which a learner might exhibit competence in a domain.

Although North American children's performance on equivalence problems in written form is nearly universally inadequate, their performance on non-symbolic problems has been more successful. The differences in performance on these diverse problem types might reflect important differences in underlying processes and/or representation (Bisanz & LeFevre, 1992). Understanding the concept of equality and what it means for two sets to be equivalent is a base skill that precedes successful performance on equivalence problems. Between the ages of 3 and 4.5 years, there is a gradual progression of being able to recognize equivalence for literal comparisons, then homogenous sets, and later heterogenous sets of objects (Mix, 1999). Acquisition of the conventional counting system enhances matching of equivalent sets of objects and may be a prerequisite for recognizing equivalence in dissimilar sets (Mix, 1999). Thus, we have some evidence of competence in these domains being related to counting among preschoolers, but counting skill is not sufficient for success on symbolic equivalence problems. This relation might lead to the suspicion that performance on symbolic equivalence problems could be related to addition and subtraction skill, but this appears not to be the case. Performance does not increase notably across the elementary school grades, while skill at addition and subtraction does (Kieran, 1981; McNeil, 2007).

Falkner et al. (1999) reported a kindergarten class that was unsuccessful with symbolic problems but successful with parallel problems modeled with blocks. Case (1985) found that 5- to 7-year old children could do physical problems with weights, and Sherman and Bisanz (2009) found that the performance of Grade 2 students was significantly better on nonsymbolic problems (presented with blocks) than on symbolic. Sherman and Bisanz (2009) also found that performance was significantly worse on part-whole (e.g., 3 + 4 = 2 + __) and combination (e.g. , 3 + 4 + 2 = 3 + __) than identity (e.g., 3 + 4 = 3 + __) and commutativity (e.g., 3 + 4 = 4 + __) problems. These discrepancies in rates of success depending on the problem presentation raise the question as to whether children are using the same underlying representations or if they perceive and represent problems differently.

Whatever the underlying representation, these patterns support the idea that the children may be able to successfully invent qualitative solutions for equivalence problems prior to being able to invent procedures for solving quantitative equivalence problems (Sherman & Bisanz, 2009). Sherman and Bisanz also found that experience solving problems in a nonsymbolic form improved performance on symbolic problems, even when the symbolic problems were presented a full week later and no explicit link was made between the two tasks.

Children also demonstrate competence in a task that appears to require very similar skills as symbolic equivalence problems. Although performance on symbolic equivalence problems is very low through Grade 6, already by Grades 2-3 children are able to choose which addend pair (e.g., 1 + 6, 3 + 5, 2 + 4) is equal to a given addend pair (e.g., 4 + 4) (Rittle-Johnson & Alibali, 1999). This task, however, does not include the equal sign, and thus this finding supports the hypothesis that a major source of children's difficulty with equivalence problems emerges from a limited understanding of the equal sign.

In a recent movement toward early algebra (EA), algebraic thought is interwoven into mathematics curricula even at the earliest grades so that children will engage in relational thinking throughout elementary and middle school (Carraher & Schliemann, 2007). Research on the best ways to implement EA is limited (Carraher & Schliemann, 2007; but see Jacobs et al., 2007), as are studies that compare various evidence-based interventions promoting EA.

The Present Study

As described above, several methods for improving children's performance on equivalence problems have been tested. Typically, however, these interventions are very short (a few minutes) and administered in one session, they are implemented one-on-one rather than in group settings, and they are compared to few if any other interventions. Thus these interventions differ markedly from children's experiences in classrooms. Furthermore, maintenance of improvements is seldom tested beyond two weeks. Finally, outcome measures vary considerably across studies, making it difficult to determine the relative effectiveness of different instructional methods. As experimental studies about cognitive development, these studies are informative and provide strong justification for believing that enhanced classroom instruction could contribute to improved outcomes, but they provide little basis for predicting the relative effectiveness of various instructional methods in the complexity of real classroom settings.

In this study assessments were conducted to determine the effectiveness of four instructional methods at improving performance on equivalence problems, whether improvement transfers to new problem types, and whether retention differs depending on instructional method. A systematic comparison of instructional methods was conducted using directly comparable instructions, procedures, and outcome measures. Children in Grades 2 and 4 participated because children in these grades typically fail equivalence problems and have shown some improvement with instruction (e.g., Carpenter et al., 2000; Jacobs, et al., 2007; McNeil, 2007; McNeil & Alibali, 2000; Perry, 1991; Rittle-Johnson & Alibali, 1999; Sherman & Bisanz, 2009; Watchorn & Bisanz, 2005). Comparisons between grades allowed for the detection of whether grade-related responses to instruction reflect improved learning or increased resistance to learning (Knuth et al., 2006; McNeil, 2007; McNeil & Alibali, 2002, 2005a,b) as a function of instructional method. Instruction took place in small groups and on two occasions, thus approximating classroom conditions while still enabling the degree of experimental control necessary for this study.

The goal of this study was to identify instructional methods that are pedagogically tractable and likely to optimize student performance. Instructional methods were designed based on an analysis of the success of the various components of the interventions described above, and also the practicality of implementation in a classroom setting. From this analysis, it was decided to exclude self-explanation (e.g., Rittle-Johnson, 2006) and comparison to other symbols (e.g., Hattikudur & Alibali, 2007) as instructional methods due to relatively limited outcomes in improving performance on equivalence problems. All conditions included pedagogically sound practices that improved performance in previous studies, including gestures (e.g., Goldin-Meadow et al., 1999), beginning each session with explaining the goal (e.g., McNeil & Alibali, 2000), and working through a minimum of five examples (e.g., Matthews & Rittle-Johnson, 2009). It was decided to focus on conceptual and generalized procedural instruction, similar to Matthews and Rittle-Johnson (2009), because of the prior success of interventions using similar methods. Manipulatives were included because of the success children have demonstrated when solving equivalence problems in nonsymbolic form (e.g., Sherman & Bisanz, 2009) and the increased success of interventions with manipulatives with younger children (e.g., Watchorn & Bisanz, 2005).

The primary focus of the study was in exploring the factors related to improving children's ability to solve equivalence problems, to transferring their learning to other types of equivalence problem, to correcting their interpretations of the equal sign, and to expanding the range of non-canonical equations they would accept as legitimate. To compare the effect of conceptual and procedural instruction and the effect of manipulatives on these outcomes, four instructional conditions were designed and the performance of students who received these conditions was compared to a control group.

The four instructional conditions were orthogonal combinations of two variables, each with two levels. One variable was the instructional focus (procedural or conceptual), and the other was manipulatives (with or without), as illustrated in Table 1. Descriptions of each of the conditions are below. These four conditions were used to examine the effect of procedural instruction in comparison to conceptual instruction, the effect of including or excluding manipulatives, and the interaction of these variables.

Table 1

Instructional Conditions

	Instructional Focus	
Instructional Materials	Procedural	Conceptual
No Manipulatives	Procedural without Manipulatives	Conceptual without Manipulatives
Manipulatives	Procedural with Manipulatives	Conceptual with Manipulatives

If manipulatives help children to connect the instruction to their existing knowledge, we would expect children in the manipulatives conditions to have superior performance to children in conditions without manipulatives. If, however, the children are only able to see the manipulatives as toys, and not as symbols relating to the equations (Uttal, Scudder, & DeLoache, 1997), the manipulatives might serve as a distraction and diminish performance among children in the manipulatives condition. Alternatively, the manipulatives and instructional focus might interact such that the manipulatives are particularly effective in one instructional condition and not the other.

Previous research shows the majority of both Grade 2 and Grade 4 students are unable to answer equivalence problems correctly (Perry, 1991; Rittle-Johnson & Alibali, 1999; Sherman & Bisanz, 2009). How the two groups might differ in their approach to the problems is unclear, as is their responsiveness to instruction. Students in Grade 4 have significantly more math skills: They have received more instruction in mathematics, have had more experience with the equal sign, and have had more opportunity to practice what they have learned. These differences may provide the Grade 4 students with an advantage. On the other hand, the Grade 4 students may have the operator view more ingrained in their thinking than Grade 2 students and could find it more difficult to start thinking of the symbol in a new way.

The direct comparison of these instructional methods in controlled settings allows conclusions about the factors involved in improving understanding about mathematical equivalence among elementary school children, and whether improvement is limited to the type of problems on which students received instruction or transfers to new problem types and other indicators of understanding. The results of this study are intended to be used in the design of materials for teachers and it is hoped that these efforts will eventually help lead to the improvement of children's performance in algebra and higher math.

This study was designed to address five main questions: (a) What are the effects of manipulatives and of procedural versus conceptual instruction in improving children's understanding of mathematical equivalence, and do these effects vary by grade and gender? (b) Does learning from instructional sessions transfer to new problem types? (c) How does instruction on equivalence problem solving affect other indicators of understanding equivalence and the equal sign? (d) What skills are associated with readiness to learn how to solve equivalence problems? and (e) Are there distinct groups of children who respond to equivalence-related tasks in qualitatively distinct ways?

Method

Design

The study consisted of four sessions, as summarized in Table 2. In all cases, children in Grades 2 and 4 were tested in groups during school hours. In Session 1, children completed tests of equivalence problem solving, problem reconstruction, equal sign definition rating, and equation rating. These paper-and-pencil tests were administered in the classroom in large groups, with each child working independently at his or her desk. The test of equivalence problem solving was used to identify children who were able to solve these problems; these children did not participate in subsequent sessions. The remaining tests in Session 1 served as pretests of the students' beliefs about the equal sign.

Table 2

Timeline Overview

Session	Activity
1	Test of equivalence problem solving
	Problem Reconstruction Test
	Definition Rating Test
	Equation Rating Test
- Two-week interval -	
2	Instruction
	Test of equivalence problem solving
- One week interval -	
3	Abbreviated test of equivalence problem solving
	Instruction
	Test of equivalence problem solving
- One month interval -	
4	Test of equivalence problem solving
	Problem Reconstruction Test
	Definition Rating Test
	Equation Rating Test

In Session 2, children who failed equivalence problems in Session 1 were randomly assigned to one of four instructional conditions (Table 1) or a control condition. Gender and school were evenly distributed across conditions. In schools with fewer than 20 students participating in a grade, only four conditions were administered. In these cases conditions were selected to equalize the total number of students in each condition across the study. Each small-group instructional session was led by one of three experimenters. Each of the experimenters conducted a similar number of sessions per condition. Following instruction, children received another test of equivalence problem solving.

Session 3 was conducted one week later and began with an abbreviated test of equivalence problem solving that consisted of the first six equivalence problems from the pretest administered in Session 1. The purpose of this test was to quickly assess how well the students maintained their learning over one week. Next the children received the same instruction as in the prior session, and then they completed another paper-and-pencil test of equivalence problem solving. In Session 4 children were retested on all of the tasks presented in Session 1 to assess whether instruction had a lasting effect.

Instruction

All children participated in one of the four instructional conditions listed in Table 1 or in a control condition. Initial group size in all instructional conditions ranged from three to six children. Sessions took approximately 20 minutes, including time to complete the posttest individually. If absences reduced the group size in Session 3 to two children, the session proceeded, but if only one child was present the session was postponed. If children asked the experimenter questions, the experimenter only repeated the instructions for that condition and explained that all questions would be fully answered at the end. All conditions began with the experimenter saying, “The goal of a problem like this is to find a number that fits in the blank so that when you put together these on this side of the equal sign (point to numbers on left), you’ll have the same number as when you put together these on this side of the equal sign (point to number and blank line on right).” The experimenter then presented the five instructional equations described below. Following instruction, children were presented a booklet with 10 practice combination equivalence problems, also described below.

Instructional Equations. In each of the four instructional conditions, the experimenter presented children with a series of five combination equivalence equations (e.g., 3 + 4 + 2 = 3 + __). Each equation was printed on a separate laminated page so that the experimenter could write on the page with a marker and erase all markings before the next instructional session. Two sets of problems were created and are presented in Table 3. In each of the instructional conditions half of the students were instructed using Set A in Session 2 and Set B in Session 3, and the other half were instructed first with Set B and then Set A.

In the conceptual conditions, one of the five examples had a large circle between the two sides of the equation instead of an equal sign, similar to one of the types of equations presented by Matthews and Rittle-Johnson (2009) in their conceptual condition. When this example was presented, the definition of the equal sign was repeated and the children were asked whether the equal sign could be placed between the two sides. The example presented in Set B displayed equivalent sides (i.e., the equal sign could be placed between the sides), whereas in Set A the two sides were not equivalent.

Table 3

Example Problems Used During Instruction

Set A	Set B
3 + 1 + 1 = 3 + ___	2 + 1 + 3 = 2 + ___
4 + 3 + 3 = 4 + ___	5 + 1 + 3 = ___ + 3
5 + 4 + 2 = ___ + 2	2 + 3 + 5 = ___ + 5
3 + 4 + 6 = ___ + 6[1]	4 + 1 + 1 = ___ + 1[1]
4 + 3 + 2 = 4 + ___	3 + 4 + 2 = 3 + ___

[1] In the conceptual conditions, this problem was presented as: 3 + 4 + 6 O 2 + 6 (Set A), and 4 + 1 + 1 O 5 + 1 (Set B), and the children were asked whether it makes sense to write an equal sign in the circle.

Practice problems. Following instruction, the children were provided with 10 practice combination equivalence problems. Multiple-choice answers were provided in the same way as for the equivalence problem solving test. See Appendix A for an example of one set of practice problems. The problems were presented in a small booklet with only one problem on each page so that children could not skip ahead of the group. Students worked independently to solve the problems. After all students had circled an answer, the experimenter provided the correct answer to the group. If a child asked why an answer was right or wrong, the same instruction presented in his or her condition was repeated.

Procedural instruction. Using appropriate gestures, students were instructed in the use of a particular algorithm (add the numbers on the left and get a sum; subtract the number on the right to get the difference; put the difference in the blank), following the protocol used by Matthews and Rittle-Johnson (2009). No further references to the concept of equality or the definition of the equal sign were made. For the example problem 2 + 1 + 3 = 2 + __, the experimenter said:

> "There's more than one way to solve this type of problem, but I'm going to show you one way to solve them today. This is what you can do: You can add the 2 and the 1 and the 3 together on the first side of the equal sign [using a marker to draw a circle around the 2 + 1 + 3] and then subtract the 2 that's over here [underlining the 2], and that amount goes in the blank. So, for this problem, think about how much 2 + 1 + 3 is? [Wait for students to think.] 2 + 1 + 3 is 6. And 6 minus 2 [pointing to the 2 on the right hand side] is 4. So our answer is 4. [Write 4 in blank.]"

The experimenter then proceeded with the four additional examples presented in Table 3, describing the procedure with each example.

Conceptual instruction. In conceptual conditions students were taught about the relational function of the equal sign and the experimenter did not discuss solution procedures, similar to Matthews and Rittle-Johnson's (2009) conceptual instruction. The experimenter began by explaining that the goal of the problem was to find a number that fit in the blank that made both sides equal.

Next, the experimenter provided the solution that would make both sides equal for the current example. The experimenter went on to explain the meaning of equal sign. For the first example in Set A, the experimenter said:

> "There are two sides to this problem: one on the left side of the equal sign [making a sweeping gesture over the left side] and one on the right side of the equal sign [making a sweeping gesture over the right side]. The first side is 3 + 1 + 1 [making a sweeping gesture]. The second side is 3 + __ [making a sweeping gesture]. What the equal sign [pointing] means is that when you put together everything on this side it's the same amount as when you put together everything on this side [sweeping hand back and forth]. Think about how much we need to put in the blank to make both sides the same. [Pause.] In this case, this side [making a sweeping gesture over the left side] is 5, so this side [making a sweeping gesture over the right side] needs to be 5 as well. To make both sides equal to 5, in this case we need to write 2 in the blank [write 2 in blank]. "

The experimenter then proceeded with the four additional examples presented in Table 3, explaining the meaning of the equal sign with each example.

Manipulatives. Children assigned to the conditions with manipulatives were presented with the same problems but with the addition of concrete materials, similar to Sherman and Bisanz (2009). As illustrated in Figure 1, a piece of blue cardboard, 6 cm high, folded to look like a tent, was placed above the equal sign and served to separate the two sides of the equations. On the child's left side of the tent there was a piece of yellow construction paper under the manipulatives and on the right there was a piece of green construction paper, to make the two sides distinctive. The corresponding written problem had a yellow background on the left and a green background on the right. Wooden cylinder blocks were placed in opaque plastic bins approximately 3 cm high to represent each term of the arithmetic expressions. Each bin was placed directly above the number it represented. For each of the five example equations, and the first three practice problems, the experimenter also displayed the equation with the blocks in the bins. The experimenter started with empty bins and a pile of blocks and asked the students to help her fill the bins to match the problem on the paper by telling her how many blocks to put in each bin. The experimenter ensured that each child had the opportunity to provide the answer on several occasions. None of the students had any difficulty with this component. Throughout the instruction an explicit effort was made to point to both the blocks and the numbers when gestures were made, to help the students make the connection between the two. For the example problems, the experimenter solved the problem by placing the appropriate number of blocks in the empty bin, and explaining to the children either the procedure (in the procedural condition) or concept (in the conceptual condition) that can be used to obtain that answer.

Figure 1. Materials used in manipulatives conditions.

Control Condition. Students in the control condition received instruction on an unrelated mathematical topic that did not involve equations and that is an activity from which children in both Grades 2 and 4 could benefit. The activity consisted of a logical reasoning task where students used number knowledge, such as place value and odd versus even numbers, to deduce the answer (e.g., "I am an odd number. You do not say me when you count by fives. I am greater than 31 but less than 36. What number am I?").

The experimenter first reviewed concepts that would be helpful for the task, such as odd and even numbers, counting by 2s, 5s, and 10s, and place value. The experimenter then guided the students through five example problems describing how to do the activity in each example. Each child then received his or her own sheet with 10 practice problems, emulating the procedures in the experimental conditions.

Testing and Scoring

Equivalence problem solving. The test of equivalence problem solving consisted of 12 problems, including four examples of each of the three types described in Table 4. These problem types have been used in previous studies (e.g., Sherman & Bisanz, 2009; Watchorn & Bisanz, 2005) and allow for comparison between learning on the problem type on which students received instruction and different problem types. Problem types were intermixed and two problem orders were created. The order of presentation was counterbalanced across children. Addends in all of the problems were single digits ranging from 2 to 9, and the sum of all digits in a single problem (i.e., the "add-all solution") ranged from 12 to 22. Children were instructed to complete as many problems as they could in a short period of time, and were asked to start at the top and go down each of two columns. Each column contained two examples of each problem type, one with the blank immediately following the equal sign and one with an addend between the equal sign and the blank. The Grade 2 students were given 8 minutes and the Grade 4 students were given 7 minutes. These time limits allowed nearly all of the students to complete all of the problems with time to spare. Scores were computed as percentage correct of the problems attempted. Each child received his or her own paper with the equations and multiple-choice answers. The multiple-choice answers were presented in an unsystematic order and included (a) the correct answer, (b) the answer that would be obtained using the add-to-equal strategy (adding all of the numbers up to the equal sign and placing the result in the blank), (c) the answer that would be obtained using the add-all strategy (adding all of the numbers present on both sides of the equal sign and placing the result in the blank), and (d) an answer that is smaller than the correct answer (to prevent children from getting credit for answering correctly if they simply select the smallest number). Children were asked to circle the answer they thought was correct. To minimize the likelihood that a child who used the add-all or add-to-equal strategy could obtain the correct answer through a minor addition or subtraction error, correct answers differed by at least two from the result that would be obtained from using either of those strategies. See Appendix B for an example of the equivalence test.

Table 4

Equivalence Problem Types

Problem Type	Example
Two-term Part-whole	a + b = c + ___
	a + b = ___ + c
Three-term Part-whole	a + b + c = d + ___
	a + b + c = ___ + d
Combination	a + b + c = a + ___
	a + b + c = ___ + a

Four options were presented for each question, and thus the chance of obtaining the correct answer by guessing was 25% for each question. With 12 questions, the probability of obtaining six or more questions correct by guessing is less than .05. Therefore, only students who were successful on five or fewer of the 12 problems (<42% correct) in Session 1 were included in the analyses.

Indicators of understanding. It is important to consider not just how students perform on the problem solving task, which they could learn to do by rote, but also to examine whether the instruction improved students' understanding of the meaning of the equal sign and how the symbol can be used. Understanding the meaning of the equal sign and appropriate problem structures was assessed with three different tasks: problem reconstruction (McNeil &

Alibali, 2002, 2004, 2005b; Rittle-Johnson & Alibali, 1999), definition rating (McNeil & Alibali, 2005a; McNeil & Alibali, 2005b; Rittle-Johnson & Alibali, 1999), and equation rating (Baroody & Ginsburg, 1983; Matthews & Rittle-Johnson, 2009; Rittle-Johnson, 2006; Rittle-Johnson & Alibali, 1999). Two versions of each task were created, with half of the students receiving Version 1 in Session 1 and Version 2 in Session 4. The remaining students received the tests in the reverse order.

Problem reconstruction task. Children were asked to reconstruct one practice and three test equations after viewing each equation for 5 seconds. All three test equations were similar to those used in the equivalence problem solving task and had operations on both sides of the equal sign. In Session 1, the problems presented were: 7 + 4 + 5 = 7 + __; 3 + 6 = 5 + __; and 4 + 1 + 6 = 6 + __. In Session 4, the problems presented were: 8 + 3 + 5 = 8 + __; 4 + 7 = 3 + __; and 6 + 2 + 5 = 5 + __. Following a coding scheme developed by McNeil and Alibali (2004), one point was awarded for each equation that was reproduced with the plus signs and equal sign in the correct locations. Numerical errors were not coded as mistakes. For example, if the equation 7 + 4 + 5 = 7 + __ was presented and the student reproduced 7 + 4 + 6 = 7 + __, the student was awarded a point for a correct reconstruction. Possible scores ranged from 0 to 3.

Definition ratings task. The definition rating task was designed to evoke participants' opinions about a number of possible definitions of the equal sign (McNeil & Alibali, 2005a). Because of time constraints, the number of definitions included in the task was reduced from six (McNeil & Alibali, 2005a) to three, but the task retained one definition from each of the categories used by McNeil and Alibali. One of the definitions corresponded to an operator view of the equal sign, another to a relational view of the sign, and one nonsense definition. Children were asked to rate the "smartness" of three fictitious students' definitions by circling an unhappy face if the definition was "not so smart," a straight face if the definition was "kind of smart," or a happy face if the definition was "very smart." See Appendix C for an example of the definitions rating task.

Points were awarded for each of the three definitions depending on the definition type. For the relational definition, children received 2 points for judging the definition as "Very Smart"; 1 point for "Kind of Smart"; and 0 points for "Not so Smart". For the operational and nonsense definitions, children received 2 points for judging the definition as "Not so Smart"; 1 point for "Kind of Smart"; and 0 points for "Very Smart". Scores were totalled so that each child received a score of 0 to 6 on the task. High scores reflect support of the relational definition of the equal sign. Internal consistency of the measure, as assessed with Cronbach's α, improved if the nonsense item was eliminated from the scale. The pattern of means across instructional conditions was similar whether the two-item or three-item scale was used and all analyses led to the same conclusions. Therefore the nonsense item was eliminated, as was done in previous studies using a similar task. Scores on the definition rating task this ranged from 0 to 4.

Equation rating task. The equation rating task was adapted from similar tasks used previously to assess children's understanding of the structure of equations (Baroody & Ginsburg, 1983; Matthews & Rittle-Johnson, 2009; Rittle-

Johnson, 2006; Rittle-Johnson & Alibali, 1999). Students indicated whether they thought each of ten equations made sense, did not make sense, or they did not know. Two examples of each of the following problem types were presented: standard (1 + 3 = 4), reverse (4 = 3 + 1), alternatives with subtraction (2 + 2 = 5 – 1), one-term identity (4 = 4); and two-term part-whole (2 + 2 = 3 + 1). Each type had one example that was correct and one example that was incorrect. See Appendix D for an example of the equation rating task.

Responses to standard problems (e.g., 1 + 3 = 4) yield no information about children's interpretation of the equal sign. These problems were included only to provide children with at least some problems that were familiar to them. Therefore responses were only analyzed for the four remaining problem types.

For each problem type, children were categorized as having a "Relational" or "Not Relational" view of the equal sign, based on their responses across both examples of that problem type. A strict criterion was set so that a child was only categorized as holding a relational view of the equal sign for that problem type if he or she responded that the correct problem "made sense" and the incorrect problem "did not make sense". All other responses were categorized as "Not relational". In separate analyses, the "Not Relational" responses were separated into "Transitional", "Operational", and "Inconsistent", but in analyzing the results the same patterns emerged as with the binary scoring. Thus scoring based on the binary coding is reported. For further details about how children were categorized in this more complex coding scheme, please refer to Appendix E. Thus, children received a score of 1 or 0 for each of the four problem types, and the total score ranged from 0 to 4.

Participants

Participants (156 from Grade 2 and 192 from Grade 4) were recruited from middle-class public schools in a suburban Canadian community. Because the focus of the study was on the effects of various forms of instruction on equivalence, only children who could not already solve equivalence problems continued beyond Session 1. Twenty-six students (7%), including 7 in Grade 2 and 19 in Grade 4, passed the pretest and so were eliminated from the study. Data from eight students were removed due to severe special needs and inability to complete the tasks, choosing to withdraw from the study, moving, or experimenter error. Data from an additional 35 students (11%) were removed because the student missed one or more of the data collection sessions. Because a minimum group size of 2 was required for the small-group instruction in most cases it was impossible to collect data at a later date from children who were absent because no other children in their condition remained untested.

Further reductions in sample size occurred because a small number children either reported that they had guessed or because the experimenter observed the child quickly and arbitrarily circling answers during the test phases of Sessions 2 or 3. The eight children in question tended to answer the practice questions either all correctly or all incorrectly. Students who answered the majority of the practice questions correctly were removed because it appeared that these students knew how to solve the problems but may have guessed on the test because they were tired of the activity. It was important not to contaminate the results by making it appear that these students did not know what they were doing when in fact they were just bored. The remaining two students solved most of the practice problems incorrectly, suggesting that they may have guessed on the test because they did not know how to solve the problems. In such cases incorrect answers accurately reflect the fact that they did not learn from the instruction. Therefore, these two students were retained in the study.

Following these adjustments, the final sample size was 273, including 122 Grade 2 students (70 girls) and 151 Grade 4 students (86 girls). Fifteen of the 75 students who were removed from the sample were never assigned to an instructional condition because they did not participate in Session 2. Of the students who had been assigned to conditions, nine were lost from each of the control, conceptual without manipulatives, and procedural without manipulatives conditions. For unknown reasons, slightly more students were lost from the conceptual with manipulatives (17) and procedural with manipulatives (16) conditions, but sample sizes remained similar across the five conditions, ranging from 52 to 56 participants. The students typically did not know the dates on which testing would occur and thus these data can be considered to be missing at random.

Results and Discussion

Reliability

To assess test-retest reliability the correlation between performance in Session 1 and Session 4 among the control group was examined for each of the measures. Correlations are presented in Table 5. Because the other conditions all received instruction that was intended to alter performance on the measures, it was only appropriate to explore test-retest reliability in the control group.

Table 5

Test-retest Correlations Among Control Group Participants and Cronbach's Alpha in Sessions 1 and 4 for All Participants

	Test-retest Reliability	Internal Consistency (Cronbach's Alpha)	
		Session 1	Session 4
Measure	n = 55	n = 253-273	n = 270-273
Equivalence Problem Solving	.25	.67	.97
Problem Reconstruction	.55***	.67	.76
Definition Rating	-.07	.43	.52
Equation Rating	.72***	.51	.70

*** $p < .001$

Test-retest reliability was low for equivalence problem solving, but in Session 1 only students who did not know how to solve equivalence problems were selected and students in the control group did not show substantial improvement. Thus the scores included in this correlation are from a highly restricted range. It is expected that guessing would be more common among students who do not know how to solve the problems, and most students in the control group did not know how to solve the problems in either Session 1 or 4. A full range of scores, including students who know how to solve equivalence problems, might have resulted in a higher correlation.

Test-retest reliability was even lower for the definition rating task. This task, however, had only two items, which may have severely limited its reliability. Also, because test-retest reliability could only be assessed with the control group, it may be affected by the same limitations as described above for the equivalence problem solving. This task, however, was meant to be equally valid for students who have operational definitions of the equal sign or relational definitions. Therefore, this low reliability indicates the measure should be interpreted with caution. The test-retest correlations for problem reconstruction and equation rating tasks were reasonably high.

The internal consistency of each of the measures in Sessions 1 and 4 was also explored, as is shown in Table 5. Internal consistencies were higher in Session 4 than Session 1 for each of the measures, but even in Session 1 the internal consistencies were reasonably high except for the definition rating task. Again, this outcome suggests that the definition rating task should be interpreted with caution.

Correlations among Measures

To determine whether the measures were tapping the same underlying construct, correlations among the measures were examined. In Session 1, performance on the equivalence problems was artificially restricted because only students who failed the task were included in the study. For this reason, equivalence problem solving was excluded from the Session 1 correlational analysis. Correlations among the three indicators of understanding are presented in Table 6.

In Session 1, the correlations were low, but the correlation between problem reconstruction and equation rating reached statistical significance. In Session 4, performance on the three types of equivalence problems was highly correlated, ranging from .91 to .93. The correlations among measures were also higher in Session 4 than in Session 1, as shown in Table 6. By Session 4 some of the students had likely benefited from instruction and had a more complete understanding of the equal sign. Thus, the greater reliability seen in Session 4 is likely due to students responding in more consistent patterns depending on their developing understanding of the equal sign and acceptable problem structures. It is interesting to note that despite the variable test-retest reliability of the measures, moderate correlations were observed in Session 4, suggesting that these tasks are tapping related skills.

Table 6

Intercorrelations Among Measures in Sessions 1 and 4

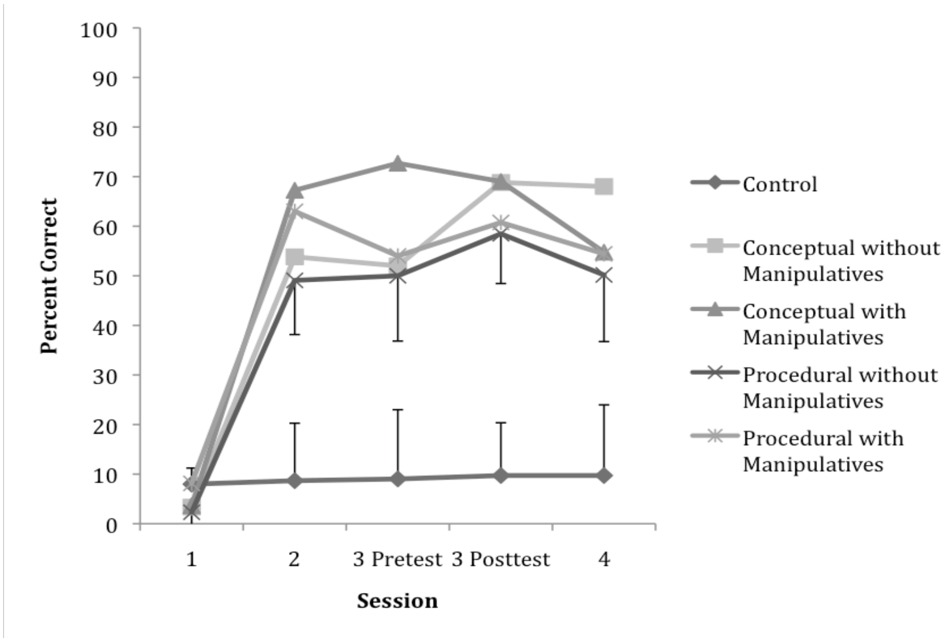
Measure	Problem Reconstruction	Definition Rating	Equation Rating
Equivalence Problem Solving	.33***	.20**	.41***
Problem Reconstruction	-	.21***	.39***
Definition Rating	.01	-	.26***
Equation Rating	.17**	.06	-

*Note. N*s 270-273. Session 1 below the diagonal, Session 4 above diagonal. Data from equivalence problem solving, Session 1, are excluded.
** $p < .01$, *** $p < .001$.

Equivalence Problem Solving

The first question that was explored was whether the children in each grade improved in their performance on equivalence problem solving as a result of instruction. Mean performance on solving equivalence problems across testing occasions is presented in Figure 2.

(a)



(b)

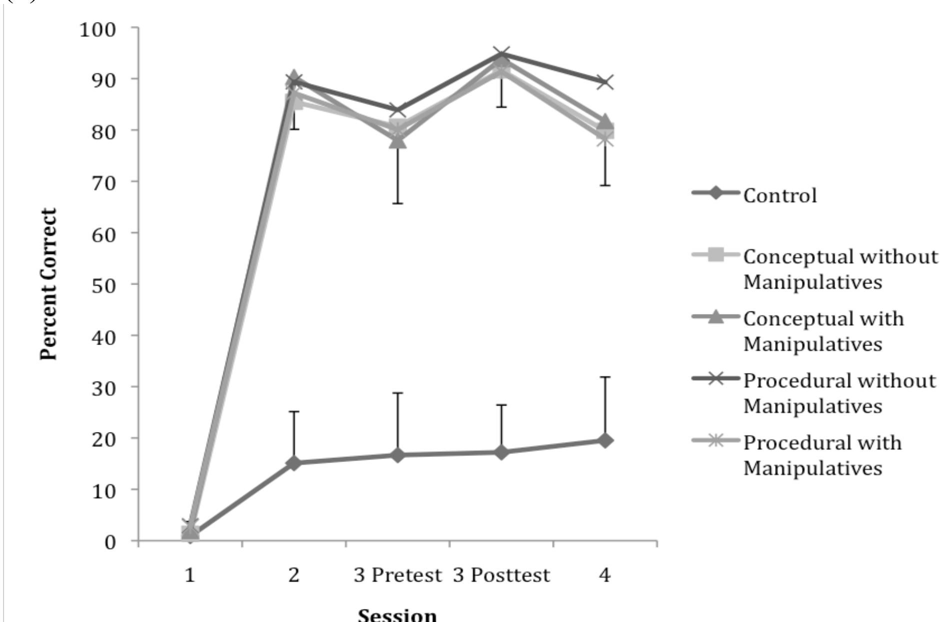


Figure 2. Performance of each instructional condition on equivalence problems across testing occasions among students in Grade 2 (Panel a) and 4 (Panel b). The upper bound of the 95% confidence interval for the control condition and the lower bound of the 95% confidence interval for the next lowest performing condition are displayed.

In Sessions 2-4 none of the 95% confidence intervals from the instructional conditions overlapped with the confidence intervals of control condition. Thus all four instructional conditions were effective. For this reason, in all subsequent analyses only the four instructional conditions on Sessions 2-4 were compared.

A primary focus of this study was identifying the factors that contributed to improving children's understanding of mathematical equivalence, and whether these factors varied by grade and gender. To determine whether the instructional effects transferred, problem type was included as a within-subjects variable. Accuracy on equivalence problems were analyzed using a 2(Grade) x 2(Gender) x 2(Instructional focus: Procedural versus Conceptual) x 2 (Manipulatives) x 3(Problem type) x 4(Test Occasion) ANOVA with repeated measures on the last two variables.

Grade 4 students (M = 85.8, SD = 26.2) outperformed Grade 2 students (M = 59.5, SD = 26.6), $F(1, 200) = 51.71$, $p < .001$, $\eta_p^2 = .205$. No effects of gender, instructional focus, or manipulatives were found (Fs < 1). Because grade interacted with problem type and gender, $F(2, 400) = 4.45$, $p = .01$, $\eta_p^2 = .022$, and with occasion, instructional focus, and manipulatives, $F(2.73, 545.85) = 2.79$, $p = .045$, $\eta_p^2 = .014$, analyses were conducted separately for each grade.

Grade 2. Two small effects were found in Grade 2. First, performance varied by problem type, $F(2, 178) = 4.07$, $p = .019$, $\eta_p^2 = .04$. Contrasts revealed that students did slightly better on combination problems (M = 61.49, SD = 30.20), on which they received instruction, than on two-term part-whole problems

(*M* = 57.38, *SD* = 34.11) and three-term part-whole problems (*M* = 59.52, *SD* = 33.09), $F(1, 178) = 6.13$, $p = .01$. However, the means only differed by a maximum of 4% and are not likely to be meaningful. Performance did not differ between and two-term and three-term part-whole problems.

The second effect was an interaction of problem type and gender, $F(2, 178) = 4.67$, $p = .011$, $\eta_p^2 = .05$, as illustrated in Figure 3. Contrasts revealed that, among boys, performance did not differ between combination problems and the two- and three-term part-whole problems, or between two-term and three-term part-whole problems. Thus, the overall effect of problem type did not apply to boys but was driven by the difference among girls, for whom performance was better on combination problems than on two- and three-term part-whole problems, $F(1, 178) = 17.20$, $p < .001$. The latter two did not differ. The reasons for this gender difference are unclear, but there is some evidence that in some circumstances boys may be better than girls at flexibly extending procedures to problem types beyond those on which they have received instruction (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Gallagher, De Lisi, Holst, McFillicuddy-De Lisi, Morely, & Cahalan, 2000).

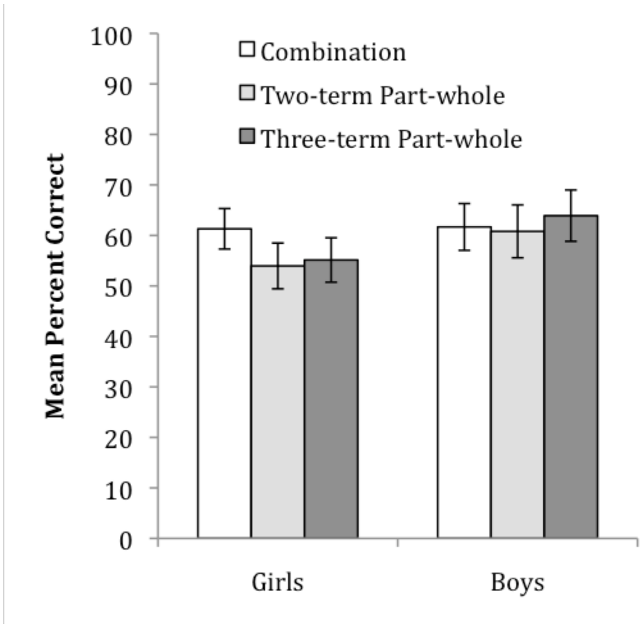


Figure 3. Mean performance of Grade 2 students as a function of problem type and gender. Bars indicate standard error.

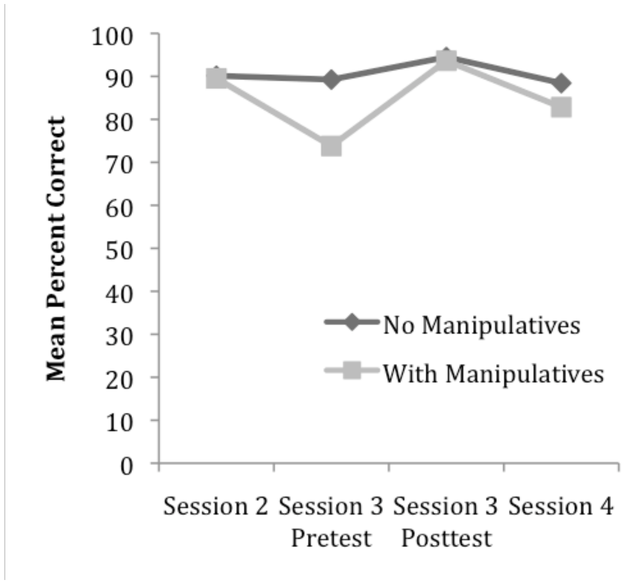
Grade 4. Three effects were found in Grade 4. First, performance differed by problem type, $F(2, 222) = 3.08$, $p = .048$, $\eta_p^2 = .03$. Orthogonal contrasts revealed that simple learning, as indexed by performance on combination problems ($M = 86.58$, $SD = 21.97$), did not differ from transfer, as indexed by the combined performance on the other two types. But a statistically significant difference existed between two-term part-whole problems ($M = 84.55$, $SD = 22.96$) and three-term part-whole problems ($M = 86.31$, $SD = 24.20$), $F(1, 178) = 4.13$, $p = .04$. It is important to note, however, that the means differed by a maximum of only 2%.

Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for test occasion ($\chi^2 = 40.09$, $p < .001$), and therefore the Greenhouse-Geisser correction was used for the next two tests.

The second effect in Grade 4 was that performance varied across the posttest occasions, $F(2.48, 274.95) = 9.27$, $p < .001$, $\eta_p^2 = .08$. Performance was highest immediately following instruction (Sessions 2 and 3 posttests) and somewhat lower following delays (Session 3 pretest and Session 4). Performance on the Session 3 posttest ($M = 92.55$, $SD = 16.74$) was higher than on the other three occasions, $F(1, 274.95) = 16.26$, $p < .001$. Performance in Session 2 ($M = 87.82$, $SD = 23.05$), which immediately followed the first instructional session, was higher than in Session 4 ($M = 81.72$, $SD = 34.19$), which was one month after the last instruction, $F(1, 274.95) = 4.99$, $p = .03$. Performance also improved between the Session 3 pretest ($M = 81.16$, $SD = 33.88$) and the Session 3 posttest, $F(1, 274.95) = 17.43$, $p < .001$.

The third effect was an interaction of manipulatives, gender, and occasion, $F(2.48, 274.95) = 3.44$, $p = .024$, $\eta_p^2 = .03$, as illustrated in Figure 4. The striking difference between genders was on the Session 3 pretest. Contrasts comparing the performance of students who received instruction with and without manipulatives revealed that among girls (Figure 4, panel a), there were no differences in Sessions 2, 3 Posttest, or 4. The same pattern was found among boys (Fs < 1.2, ps > .28). On the Session 3 pretest, however, the pattern was different among boys than among girls. Among boys, the students who had received instruction without manipulatives had lower performance than the students who had received instruction with manipulatives, $F(1, 274.95) = 5.48$, $p = .02$. Among girls, the reverse was true, $F(1, 274.95) = 8.90$, $p < .01$. Specifically, the drop in performance on the Session 3 pretest was observed among girls who received instruction with manipulatives, whereas among boys, the drop was observed in the group that did not receive instruction with manipulatives. The effect size for this interaction was small and there is no theoretical reasoning for the observed pattern, and so this interaction was not interpreted further.

(a)



(b)

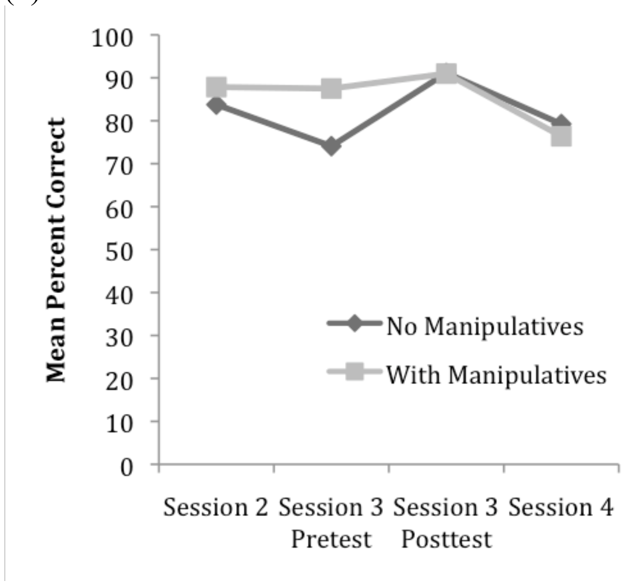


Figure 4. Effect of manipulatives on the mean percent correct on equivalence problems for Grade 4 girls (Panel a) and boys (Panel b) across occasions.

Summary. Grade 4 students outperformed Grade 2 students, but all four instructional conditions were effective at improving performance on equivalence problems in both grades. Small effects were found within each grade, but none involved instructional focus.

Indicators of Understanding

For ease of comparison between tasks, the same types of analysis are reported for the three indicators of understanding. Analyses could be done in several ways, and each option presented advantages and disadvantages. Parametric testing (e.g., ANOVA) is typically powerful, provides interpretable results, allows controls for multiple comparisons, and allows for tests of interactions. However, there were obstacles to using ANOVA, for example, scores on the problem reconstruction task were not normally distributed, violating one of the assumptions of ANOVA. To solve this problem one option is to compute difference scores between performance on Session 4 and Session 1, but the reliability of difference scores is low, especially when using measures with low reliability.

An alternative is nonparametric testing. Nonparametric tests can be used with data that are not normally distributed and for counts in categories as opposed to scores, but are typically less powerful and do not allow for testing interactions involving repeated measures.

With these considerations in mind, a variety of parametric and nonparametric analyses were conducted. The results of the parametric analyses are presented below, but were generally confirmed with nonparametric analyses. For each indicator, a one-way ANOVA with condition as the between subject variable was conducted on Session 1 scores to check for pre-existing differences between groups. Next, to test for improvement between Session 1 and Session 4 a similar ANOVA was conducted on the difference scores with condition as a between subjects variable. The intercept was examined to assess overall improvement, and 95% confidence intervals were inspected for each condition. Finally, to explore the improvements, a Grade x Gender x Instructional Focus x Manipulatives ANCOVA on Session 4 scores was conducted, with the corresponding Session 1 score as the covariate.

Problem reconstruction. Performance on the problem reconstruction task was low across both sessions, as illustrated in Figure 5.

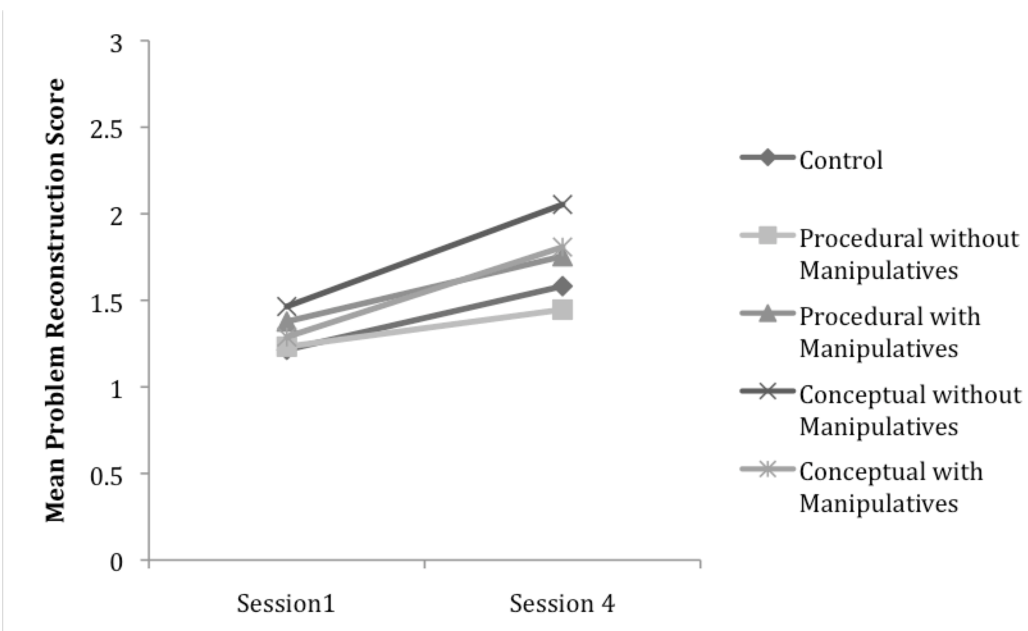


Figure 5. Mean points (maximum of 3) on problem reconstruction task in Sessions 1 and 4 by instructional condition.

In Session 1, none of the conditions differed from each other. An overall improvement on problem reconstruction scores was observed between Sessions 1 and 4, $F(1, 267) = 34.61$, $p < .001$, $\eta_p^2 = .12$. The 95% confidence intervals of the mean difference scores overlapped with zero for the group that received procedural instruction without manipulatives, indicating that this group did not improve. All other groups improved.

To explore these improvements observed in Session 4, a 2(Grade) x 2(Gender) x 2(Instructional focus: procedural versus conceptual) x 2(Manipulatives) ANCOVA on the Session 4 problem reconstruction scores was conducted, with the Session 1 problem reconstruction score as the covariate. Grade 4 students (adjusted $M = 1.95$, $SD = 1.10$) outperformed Grade 2 students (adjusted $M = 1.57$, $SD = 1.13$), $F(1, 200) = 5.72$, $p = .018$, $\eta_p^2 = .03$, and the students who received conceptual instruction (adjusted $M = 1.90$, $SD = 1.07$) outperformed students who received procedural instruction (adjusted $M = 1.62$, $SD = 1.04$) $F(1, 200) = 3.91$, $p = .049$, $\eta_p^2 = .02$. Gender and manipulatives did not affect scores, and no significant interactions were found.

Because a primary interest in this study was the effect of instructional focus, it was important to compare the procedural and conceptual groups to the control group. In the 4-way ANCOVA described above, the control group was excluded to test the effect of the instructional focus and manipulatives variables. To compare the conceptual and procedural groups to the control group, contrasts were conducted using the adjusted means from a 2(Grade) x 2(Gender) x

5(Condition) ANCOVA on the Session 4 problem reconstruction scores, with the Session 1 problem reconstruction score as the covariate.

The advantage of conceptual instruction over the control group was not statistically significant, *p* = .08, but the difference was in the expected direction, as illustrated in Figure 6. Students who received procedural instruction did not differ from the control group. The ability to reconstruct equivalence equations accurately is associated with a relational view of the equal sign. Thus, it appears that the relation between conceptual instruction and a relational understanding of the equal sign is worthy of future investigation.

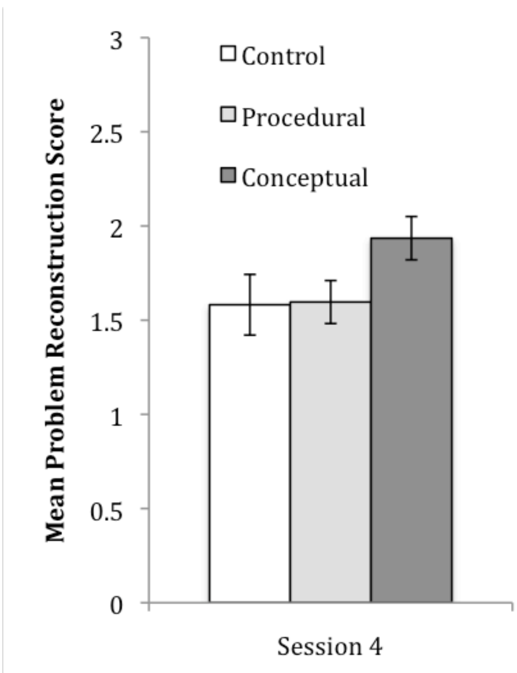


Figure 6. Mean score (maximum of 3) on problem reconstruction task for each instructional focus group in Session 4. Bars indicate standard error.

Definition rating. Performance on the definition rating task is illustrated in Figure 7.

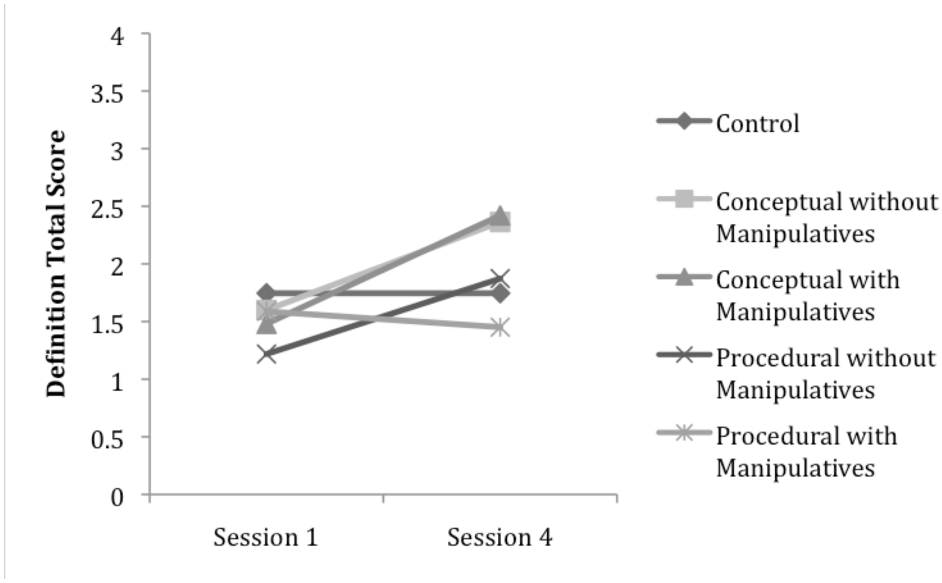


Figure 7. Mean score (maximum of 4) on definition rating task in Sessions 1 and 4 for each instructional condition.

In Session 1, none of the conditions differed. An overall improvement between Sessions 1 and 4 was observed, $F(1, 261) = 15.95$, $p < .001$, $\eta_p^2 = .06$. The 95% confidence intervals of the mean difference scores for the control condition and the procedural with manipulatives condition overlapped with zero, indicating these two groups did not improve between Sessions 1 and 4. The remaining three groups all improved.

To explore these improvements a 2(Grade) x 2(Gender) x 2(Instructional focus: procedural versus conceptual) x 2(Manipulatives) ANCOVA on the Session 4 definition rating scores was conducted, with the Session 1 definition rating score as the covariate. Grade 4 students (adjusted $M = 2.36$, $SD = 1.25$) outperformed Grade 2 students (adjusted $M = 1.57$, $SD = 1.25$), $F(1, 194) = 20.74$, $p < .001$, $\eta_p^2 = .10$, and the students who received conceptual instruction (adjusted $M = 2.32$, $SD = 1.27$) outperformed students who received procedural instruction (adjusted $M = 1.62$, $SD = 1.25$), $F(1, 194) = 16.21$, $p < .001$, $\eta_p^2 = .08$. Gender and manipulatives did not affect scores, but the effect of instructional focus was not consistent across grade and gender, $F(1, 194) = 4.06$, $p = .045$, $\eta_p^2 = .02$, as illustrated in Figure 8.

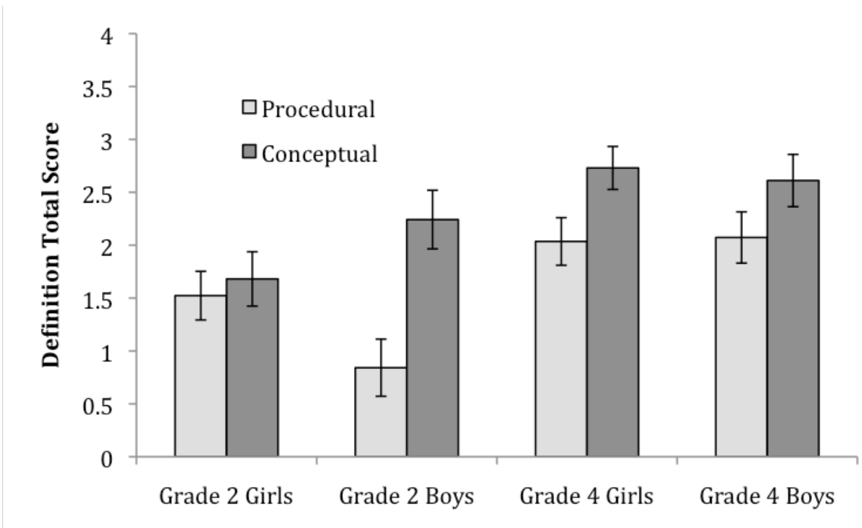


Figure 8. Effect of instructional focus on definition rating total score (out of 4) for each grade and gender. Bars indicate standard error.

To explore this interaction, separate ANCOVAs were conducted for each grade. In Grade 4, students in the conceptual condition outperformed those in the procedural condition, $F(1, 109) = 7.58$, $p = .007$, $\eta_p^2 = .07$, but in Grade 2 the effect varied by gender. Grade 2 boys who received conceptual instruction outperformed those with procedural instruction, $F(1, 36) = 13.69$, $p = .001$, $\eta_p^2 = .28$, but the two groups did not differ among the girls.

To further explore the effect of instructional focus, contrasts were conducted using the adjusted means from a 2(Grade) x 2(Gender) x 5(Condition)

ANCOVA on the Session 4 definition rating score, with the Session 1 definition rating score as the covariate. Children in the conceptual group outperformed students in the control group, $p = .003$, as illustrated in Figure 9, but there was no difference between the procedural group and the control group. This result suggests that conceptual instruction may help students recognize a more accurate interpretation of the meaning of the equal sign. It is interesting to note that the expected pattern was found despite the low reliability of this measure.

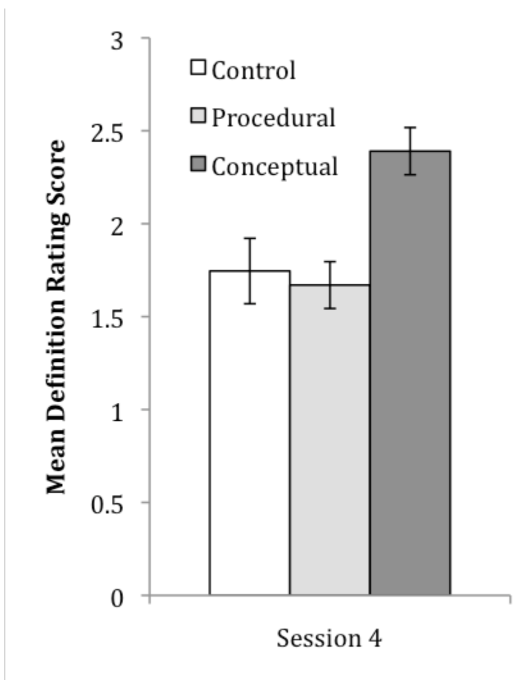


Figure 9. Mean score (maximum of 4) on definition rating task for each instructional focus group in Session 4. Bars indicate standard error.

Equation rating. For each problem type on the equation rating task children were categorized as have responded in a way consistent with a relational understanding of the equal sign or not. Scores of 1 or 0 for each of the four problem types were added to compute an equation rating total score out of 4.

In Session 1, none of the groups differed. An overall improvement between Sessions 1 and 4 was observed, $F(1, 268) = 53.05$, $p < .001$, $\eta_p^2 = .17$. The 95% confidence intervals of the mean difference scores for the control condition overlapped with zero, indicating that the control group did not improve between Session 1 and 4. All four instructional conditions improved.

Performance on the equation rating task was generally low in that the mean score was below 2 (out of a possible maximum of 4) for all of the instructional conditions, even in Session 4, as illustrated in Figure 10.

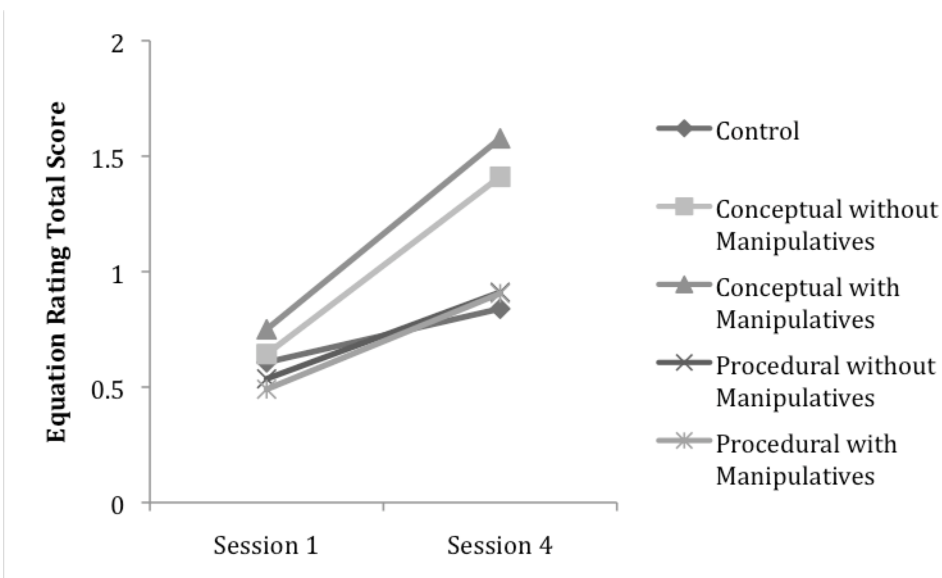


Figure 10. Mean equation rating score (maximum of 4) for each instructional condition in Sessions 1 and 4.

To explore the instructional groups' improvements a 2(Grade) x 2(Gender) x 2(Instructional focus: procedural versus conceptual) x 2(Manipulatives) ANCOVA on the Session 4 total scores was conducted, with the Session 1 total score as the covariate.

Grade 4 students (adjusted $M = 1.47$, $SD = 1.22$) outperformed Grade 2 students (adjusted $M = 0.94$, $SD = 1.22$; $F(1, 200) = 9.89$, $p = .002$, $\eta_p^2 = .05$) and boys (adjusted $M = 1.42$, $SD = 1.22$) outperformed girls (adjusted $M = 0.99$, $SD = 1.21$; $F(1, 200) = 6.64$, $p = .01$, $\eta_p^2 = .03$). Consistent with the other two indicators of understanding, students who received conceptual instruction (adjusted $M = 1.46$, $SD = 1.24$) outperformed those who received procedural instruction, (adjusted $M = 0.96$, $SD = 1.21$; $F(1, 200) = 8.98$, $p =.003$, $\eta_p^2 = .04$). The use of manipulatives did not affect scores.

Next, the conceptual and procedural groups were compared to the control group. Contrasts were conducted using the adjusted means from a 2(Grade) x 2(Gender) x 5(Condition) ANCOVA on the Session 4 equation rating score, with the Session 1 equation rating score as the covariate. The conceptual group exceeded the control group, $p < .001$, as illustrated in Figure 11, and the procedural and control groups did not differ, $p = .48$. This result suggests that the conceptual instruction helped students accept non-canonical equations.

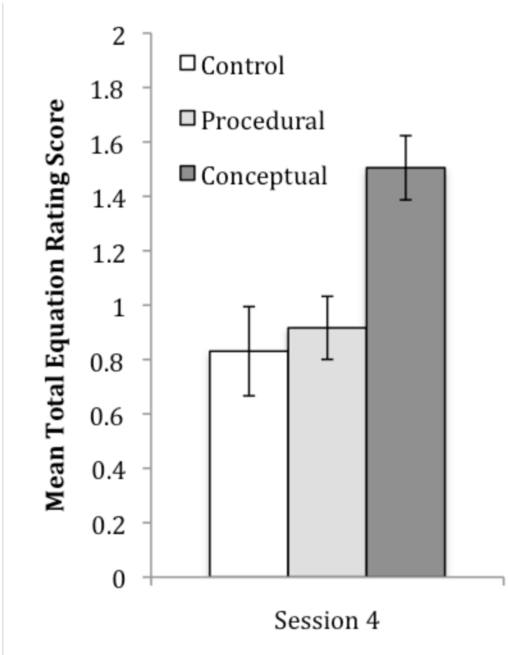
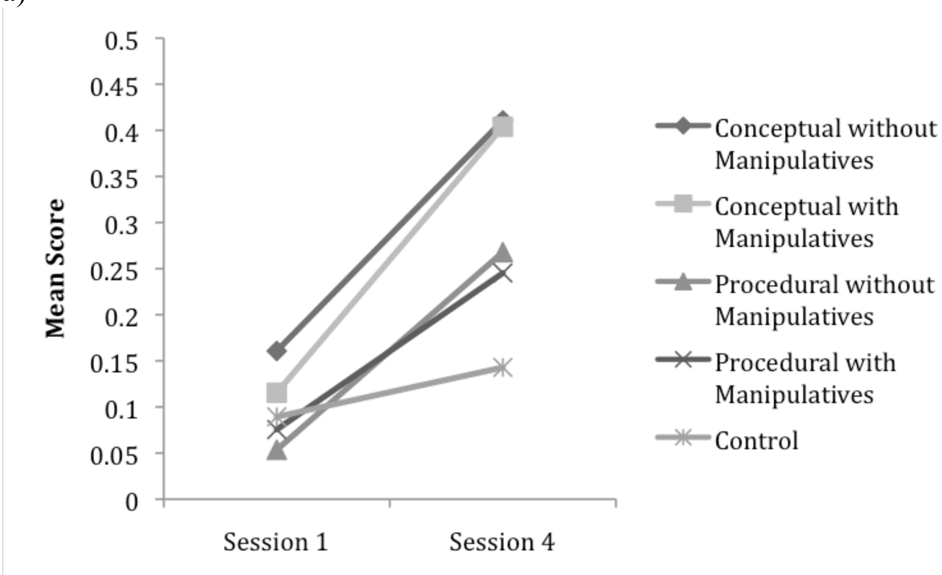


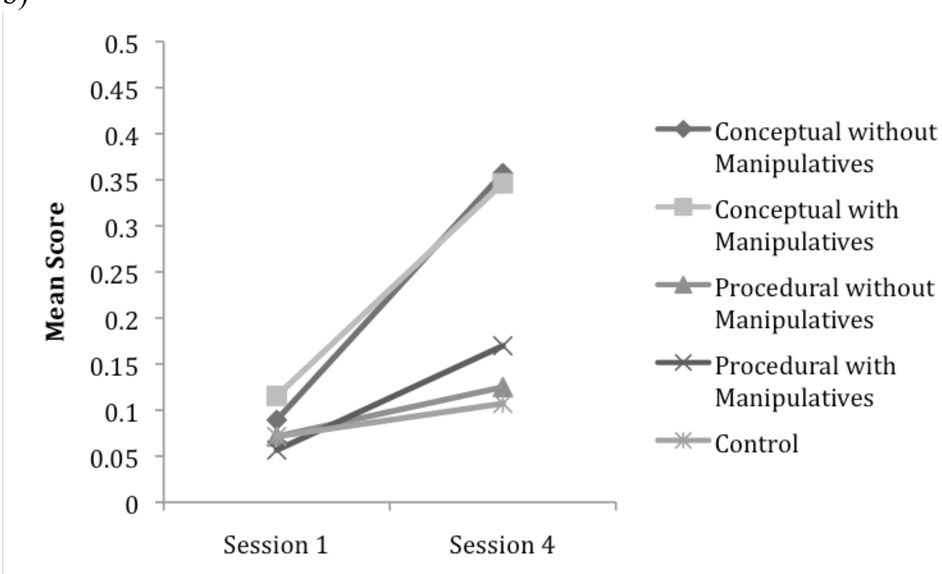
Figure 11. Performance on equation rating task (out of 4) for each instructional focus group in Session 4. Bars indicate standard error.

Equation rating by problem type. Performances for each condition on each equation rating problem type in Sessions 1 and 4 are presented in Figure 12. To explore whether the patterns found overall on the equation rating task differed by problem type, a 2(Grade) x 2(Gender) x 2(Instructional focus: procedural versus conceptual) x 2(Manipulatives) ANCOVA on the Session 4 equation rating score for each problem type was conducted with the Session 1 equation rating score as the covariate. Because of the interest in the effect of instructional focus, for each significant instructional focus effect the conceptual and procedural groups were compared to the control group. To compute the appropriate adjusted means for each contrast, a 2(Grade) x 2(Gender) X 5(Condition) ANCOVA was conducted on the Session 4 equation rating score for that problem type, with the Session 1equation rating score for that problem type as the covariate.

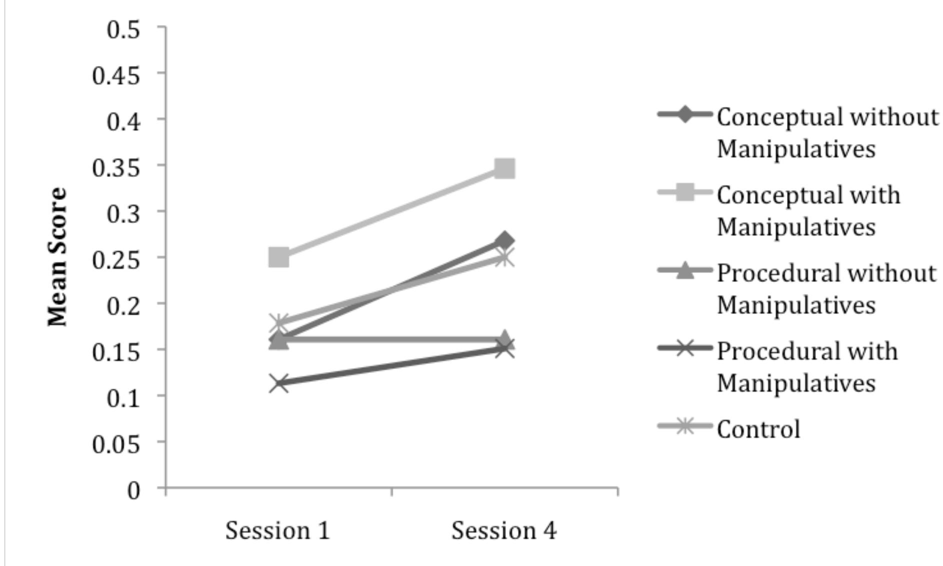
a)



b)



c)



d)

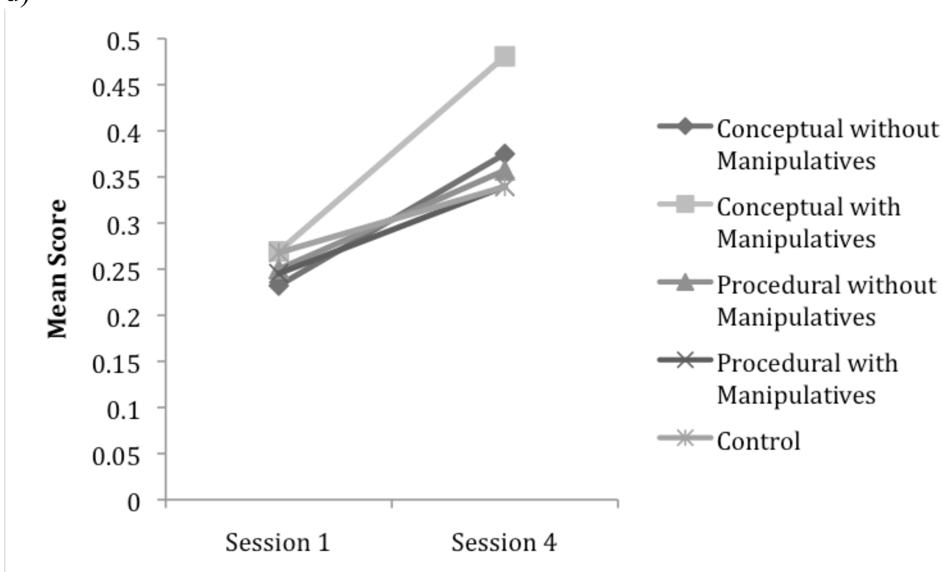


Figure 12. Means (maximum of 1) for each instructional condition on the equation rating task on part-whole (Panel a), alternatives with subtraction (Panel b), one-term identity (Panel c), and reverse problems (Panel d).

Part-whole problem type. Grade 4 students (adjusted $M = .43$, $SD = 0.47$) outperformed Grade 2 students (adjusted $M = .22$, $SD = 0.47$), $F(1, 200) = 10.73$,

p = .001, η_p^2 = .05. The difference between students who received conceptual instruction (adjusted M = 0.39, SD = 0.48) and students who received procedural instruction (adjusted M = 0.27, SD = 0.47) did not reach statistical significance but was suggestive, $F(1, 200)$ = 3.66, p = .057, η_p^2 = .02. Contrasts revealed that the conceptual group outperformed the control group, p < .001, but the procedural group did not, as illustrated in Figure 13.

Alternatives with subtraction problem type. Students who received conceptual instruction (adjusted M = 0.35, SD = 0.44) outperformed students who received procedural instruction (adjusted M = 0.15, SD = 0.43), $F(1, 200)$ = 12.08, p = .001, η_p^2 = .06. Contrasts confirmed that the conceptual group outperformed the control group, p < .001, but the procedural group did not (see Figure 13).

One-term identity problem type. Students who received conceptual instruction (adjusted M = 0.31, SD = 0.39) outperformed students who received procedural instruction (adjusted M = 0.18, SD = 0.39), $F(1, 200)$ = 6.72, p = .010, η_p^2 = .03. Boys (adjusted M = 0.34, SD = 0.39) also outperformed girls (adjusted M = 0.15, SD = 0.39), $F(1, 200)$ = 13.19, p < .001, η_p^2 = .06. Contrasts comparing the conceptual and procedural groups to the control group did not reveal any differences (see Figure 13).

Reverse problem type. Grade 4 students (adjusted M = .54, SD = 0.47) outperformed Grade 2 students (adjusted M = .23, SD = 0.48), $F(1, 200)$ = 22.40, p < .001, η_p^2 = .10, and boys (adjusted M = 0.45, SD = 0.46) also outperformed girls (adjusted M = 0.32, SD = 0.46), $F(1, 200)$ = 4.33, p = .039, η_p^2 = .02. No other effects were found.

The conceptual group's performance was still the highest of the groups on the reverse problems, but the control and procedural groups also performed relatively well on this problem type, as illustrated in Figure 13. Thus, for three of the four problem types, students who received conceptual instruction were more likely to rate equations in a pattern consistent with a relational view of the equal sign.

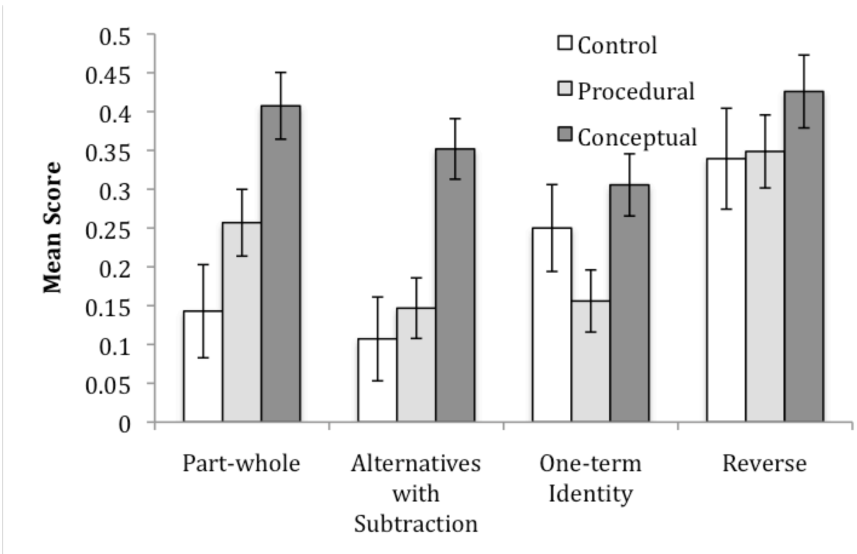


Figure 13. Means (maximum of 1) in Session 4 reflecting the percent of students responding in a pattern consistent with a relational understanding of the equal sign on each problem type, by instructional focus on the equation rating task. Bars indicate standard error.

Summary of indicators of understanding. The influence of instruction was generally consistent across the indicators of understanding. Students who received conceptual instruction generally outperformed students who received procedural instruction on problem reconstruction, definition rating, and equation rating. The use of manipulatives during instruction did not affect performance.

Individual Differences

The Effect of Pre-existing Skill on Learning. This study also afforded the opportunity to address the question of how children's pre-existing skill, as measured by the indicators of understanding in Session 1, might facilitate learning, as demonstrated by performance on equivalence problems in Session 4. A regression was conducted with grade, gender, problem reconstruction score, definition rating score, and equation rating scores from Session 1 as independent variables, and total percent correct on the equivalence problems in Session 4 as the dependent variable. Students in the control condition were excluded because they did not receive instruction on equivalence problems.

Problem reconstruction and equation rating in Session 1 uniquely accounted for a significant amount of the variability in equivalence problem scores in Session 4, as indicated in Table 7. In separate regressions for Grades 2 and 4, the same pattern occurred but equation rating failed to reach significance in Grade 4. Thus skills associated with performance on the problem reconstruction task, which may include mental models of equation structures and attentional skill, are associated with learning to solve equivalence problems when provided with instruction. The children who are initially better at problem reconstruction may have the cognitive structures in place that allow them to take advantage of the instruction.

Table 7

Squared Semi-partial Correlation Coefficients of Session 1 Measures Associated with Performance on Session 4 Equivalence Problems

Predictor	Overall	Grade 2	Grade 4
Grade	.03*	/	/
Gender	.00	.01	.01
Problem Reconstruction score	.06***	.10**	.05*
Definition Rating score	.00	.00	.01
Equation Rating	.03**	.04*	.02
R^2	.211***	.175***	.098**

* $p < .05$, ** $p < .01$, *** $p < .001$

Profiles of Performance. If groups of individuals are identified by characteristic skill profiles, we may see relations between these skills that are not captured in correlational analyses. A cluster analysis was used to explore how performance on equivalence problem solving, problem reconstruction, definition rating, and equation rating clustered in Session 4. Children from all five instructional conditions were included in the analysis, but because of missing data the sample size was reduced to 270.

Ward's (1963) hierarchical agglomerative method was first employed to investigate potential solutions. A large increase in fusion coefficients indicates that the clusters being merged are dissimilar. All increases in fusion coefficients were quite small until four clusters were merged into three (583.37 to 771.11). A second reasonably large increase in fusion coefficients occurred when three clusters were merged into two (771.11 to 1045.26). When two clusters were merged into one a much larger jump (from 1045.26 to 1492.90) occurred, reflecting in a substantial loss of information and resulting in a single cluster that would be uninformative. The one-cluster solution was therefore discounted. Thus examination of two-, three-, and four-cluster solutions was warranted.

Using squared Euclidean distance seed values from Ward's (1963) method, a *k*-means cluster analysis was conducted based on each of the possible cluster solutions (k = 4, 3, and 2 means). *K*-means cluster analysis is an iterative partitioning method where seed values estimate the initial centroids of the *k* clusters, but as each member is assigned to a cluster the centroid is recalculated. Iterations through the dataset continue re-assigning each case to its nearest cluster centroid until no further iterations can provide a more optimal assignment of cases to clusters (Aldenderfer & Blashfield, 1984).

For ease of interpretation, all scores were converted into percentages of the total possible score for that measure. Recall that equivalence problem solving was a multiple-choice test and thus only scores equal to or greater than 50% were interpreted as reflecting an ability to solve equivalence problems. In Figure 14 the four-cluster solution is displayed. One group of children performed poorly on all measures. Because the tasks are meant to tap whether the children have an operational or relational definition of the equal sign and low scores reflect an operational understanding of the equal sign, the first group can be called

"Operational Thinkers" (n = 59). The next group had high scores on the definition rating task, but low scores on the three other tasks. This group was entitled the "Good Definers" (n = 62). A third group had moderate performance on equivalence problem solving, high performance on problem reconstruction and definition rating, but low performance on the equation rating task. This group was entitled the "Poor Equation Raters" (n = 85). A final group, the "Relational Thinkers," had high performance on all four tasks (n = 64).

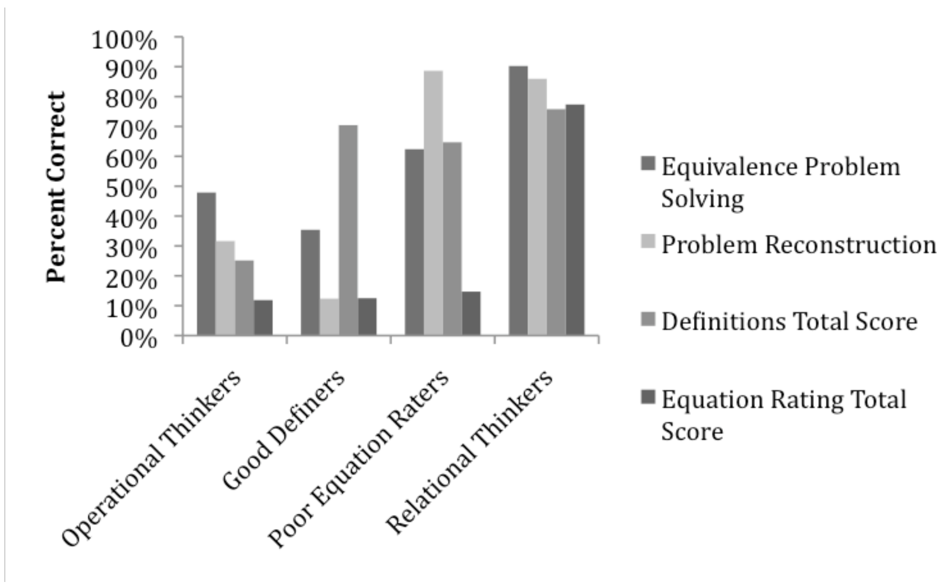


Figure 14. Four cluster solution.

Next, the three-cluster solution was examined, as illustrated in Figure 15.

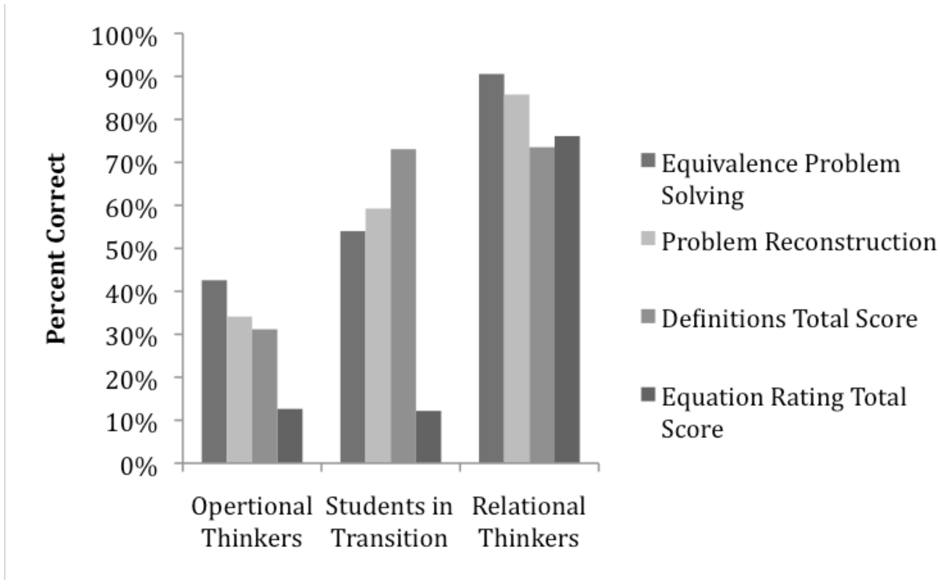


Figure 15. Three cluster solution.

The three-cluster solution essentially collapsed the two middle groups from the four-cluster solution and created an intermediate group, entitled the "Students in Transition" (n = 117) for the purposes of this study. The "Operational Thinkers" (n = 85) remained on the low end and the "Relational Thinkers" (n = 68) on the high end.

The two-cluster solution simply divided the students into high (n = 103) and low (n = 167) performance groups, as illustrated in Figure 16.

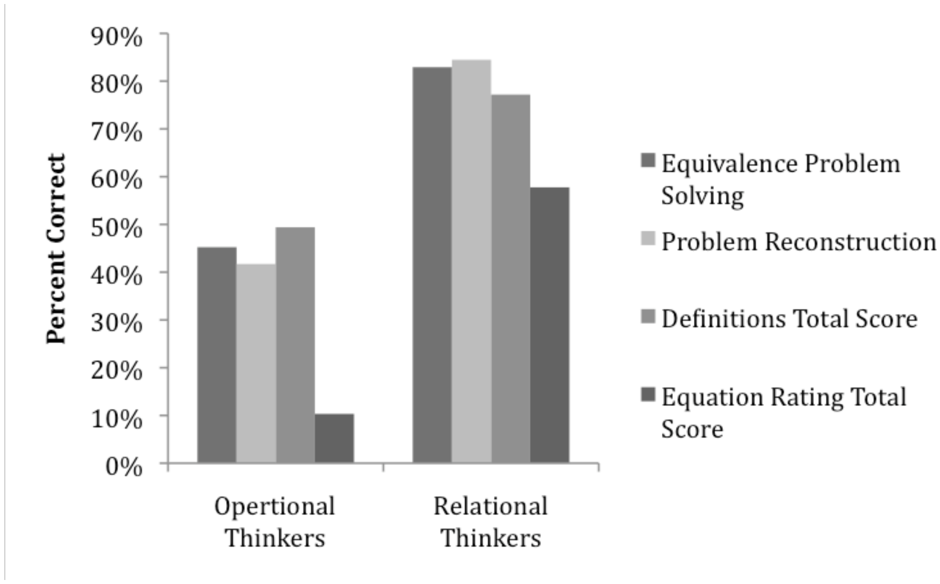


Figure 16. Two cluster solution.

Because each of the clusters in the four-cluster solution were substantial in size and potentially important information was lost in the collapsing into fewer clusters, the four-cluster solution was retained.

The strengths in the four clusters appear to be additive, in that the clusters could be arranged in an order such that once a strength appeared, other strengths were only added to it; no strengths disappeared. For example, the students in the "Operational Thinkers" cluster do not perform well on the definition rating task. High scores on this task first appear in the "Good Definers" cluster, and are also found in the next two clusters (i.e., "Poor Equation Raters" and "Relational Thinkers"). Similarly, high scores on the problem reconstruction task first appear in the "Poor Equation Raters" cluster and are maintained in the "Relational Thinkers" cluster. This observation led to a hypothesis about the progression of developing a relational understanding of the equal sign. Children may first acquire a relational definition of the equal sign, but only in a limited or superficial way. They are unable to apply this definition to understand non-canonical equations. Next, children start to build a mental structure that accommodates non-canonical equations, which facilitates performance on reconstructing non-canonical equations. The new mental structure, however, is still limited in that it does not accommodate all types of non-canonical equations, and performance remains low on equation rating. Lastly, children are able to truly generalize their new definition of the equal sign and performance is high on all tasks.

If the proposed progression exists and children in Grade 4 benefited from instruction more than children in Grade 2, we would expect to see more children in Grade 2 at the lower end of the progression and more children in Grade 4 at the upper end. Table 8 displays the percentage of students within each Grade in each of the four clusters.

Table 8

Percentage of Students in Each Cluster by Grade

Grade	Operational Thinkers	Good Definers	Poor Equation Raters	Relational Thinkers	Total
2	33.6%	29.4%	23.5%	13.4%	100%
4	12.6%	17.9%	37.7%	31.8%	100%

In Grade 2, the percentage of students in each cluster declines through the proposed progression. In Grade 4, the percentage increases across the first three, followed by a slight decline in the percentage of students in the fourth and final cluster. This pattern is consistent with the proposed progression, where Grade 2 students are typically less advanced along the progression than Grade 4 students. This progression is speculative and may be an artifact of the specific instruction that was provided, but it serves as a hypothesis that can be tested further in longitudinal studies.

Lastly, the effect of instructional focus on cluster placement was explored. As is demonstrated in Figure 17, approximately one-third of the children were classified as Operational Thinkers, regardless of instructional focus. The percentage of students from each instructional focus in the other clusters varied, but the conceptual group had the highest percentage in the Relational Thinkers cluster.

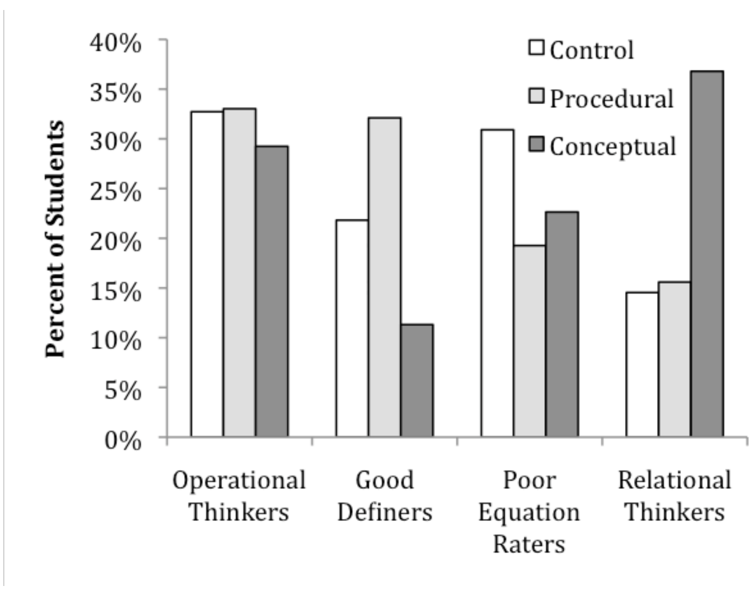


Figure 17. Percentage of students within each instructional focus condition who were in each of the four clusters.

Summary. Pre-existing skill at problem reconstruction and equation rating was associated with post-instruction performance on the equivalence problem-solving task. This finding suggests that understanding problem structure and attentional skill may be associated with the ability to benefit from instruction on equivalence problems.

Individual differences in performance profiles were found such that children clustered into four groups based on their performance across tasks relating to mathematical equivalence. Examination of these clusters led to a hypothesis about the development of children's understanding of the equal sign. The pattern suggests that when children first acquire a relational definition of the equal sign, it is only in a limited or superficial way. Children gradually learn to apply this new definition to understand non-canonical equations, however not all non-canonical equation types are accepted at the same time. Lastly, conceptual instruction appears to facilitate progression to a more advanced understanding of the equal sign, as the group that received conceptual instruction had the highest percentage of students in the Relational Thinkers cluster.

General Discussion

For generations children have struggled to overcome an operator interpretation of the equal sign (e.g., Behr et al., 1976; Goldin-Meadow et al., 1999; Renwick, 1932). The problem is so pervasive in Canada and the United States that some have even hypothesized that intellectual development or general conceptual limitations may be hindering students ability to view the equal sign in a relational way (Denmark et al. 1976). Performance in other countries, however, has provided evidence that children in Grades 2 through 6 are capable of succeeding on equivalence problems (e.g., Watchorn et al., 2009).

Correcting children's misconceptions is not an easy task. Most interventions, although often producing improvements that are statistically significant, have had only limited success, putting into question the meaningfulness of the improvements in the classroom setting (e.g., Alibali, 1999; McNeil, 2008). Moreover, it is difficult to directly compare the interventions' relative effectiveness because the outcome measures varied considerably across studies. The instructional methods that appear to have been most successful involve extensive teacher training (Carpenter et al., 2000), which requires investments of time and resources beyond the capacities of many schools.

A review of the existing literature led to the question of whether it is possible to improve student learning about equivalence in a meaningful way using a teacher-directed approach. Because of the importance of procedural and conceptual instruction in mathematics, both were compared to explore which is most effective at improving student performance. Several additional variables were also of interest. Children have had more success on problems in nonsymbolic form (Sherman & Bisanz, 2009), so manipulatives were included in instruction to determine whether they affect improvements. In previous research 9-year-olds demonstrated increased resistance to learning in comparison to 7-year-olds (McNeil, 2007), and thus the effect of instruction was compared between these two age groups.

The importance of examining the different contexts within which a learner might exhibit competence in a domain has recently been highlighted by Schneider and Stern (2010). Children's ability to transfer their learning varies between children and depending on the instruction they received. To conclude that the instruction led to meaningful improvements, it was important to examine how children performed on new problem types that differed from the ones on which they were instructed and on other indicators of understanding. Procedural instruction has been detrimental in transfer tasks in at least one study (Perry, 1991), and even with the most successful interventions the difficulty of changing students definitions of the equal sign has been noted (Carpenter et al., 2003).

Few researchers have examined the relation between children's pre-existing skills and subsequent performance on equivalence problems. McNeil and Alibali (2005b) noted the relation between knowledge of which feature of problems to encode and subsequent learning on equivalence problems, and this study afforded the opportunity to confirm this relation and search for other relations. Further, no one has examined post-instruction patterns of performance profiles across equivalence tasks through a cluster analysis.

The purpose of this study was to compare pedagogically tractable instructional methods designed to optimize student performance on equivalence measures. Two variables were explored—procedural or conceptual instructional focus, and the use of manipulatives—and four instructional conditions were created representing the orthogonal combinations of these two variables. A systematic comparison of the instructional conditions was conducted using directly comparable instructions, procedures, and outcome measures among students in Grades 2 and 4. This approach was used to investigate improvements in solving equivalence problems, whether improvements transferred to new problem types, whether their retention differed depending on instructional method, and whether improvements were seen on other indicators of understanding.

Five main questions were addressed. First, what are the effects of manipulatives, procedural instruction, and conceptual instruction in improving children's performance on solving equivalence problems, and do these effects vary by grade and gender? A systematic comparison of instructional methods was conducted using directly comparable instructions, procedures, and outcome measures. All instructional conditions improved problem solving, and the improvements were substantial. Students in Grade 2 approached and students in Grade 4 matched the high level of performance previously observed in Asian countries (e.g., Watchorn et al. 2009). These significant improvements are also similar to the improvements seen with approaches that involve intense teacher training such as Cognitively Guided Instruction (CGI). The improvements were also maintained at a reasonably high level for one month after instruction, longer than any retention period to be tested with equivalence problems in the past.

Grade 4 students outperformed Grade 2 students, reflecting improved learning in the older students as opposed to the increased resistance to learning that has been observed in some other studies (e.g., McNeil, 2007; McNeil & Alibali, 2002, 2005a,b). This finding implies that students who have had more experience with canonical addition and subtraction are not necessarily more entrenched in their operational view of the equal sign.

The second question was whether learning from the instructional sessions transferred to new problem types. If benefits were only seen on the same types of problems as those on which students received instruction, it would indicate a severe limitation on the intervention. Instruction was provided only on combination equivalence problems, and proficiency was later tested on both combination and part-whole problems. Learning transferred to the part-whole problem types, but boys were able to transfer their learning somewhat better than girls. The effect size for this gender difference was small, but the pattern is consistent with the findings in previous studies that boys may be better than girls at flexibly extending procedures to problem types beyond those on which they have received instruction (e.g., Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Gallagher, De Lisi, Holst, McFillicuddy-De Lisi, Morely, & Cahalan, 2000).

In contrast to Perry (1991), students who received procedural instruction did not suffer in their performance on problems that differed from the problem type on which they received instruction. It is important to note, however, that some of the transfer problems used in Perry's study were multiplication problems, which were not assessed in this study.

The third question addressed was how instruction on equivalence problem solving affects other indicators of understanding. If children have improved in their ability to solve only equivalence problems after instruction, we must question whether they understand what they are doing. Students who have memorized how to solve equivalence problems but maintain an operational interpretation of the equal sign may have difficulty transitioning to algebra, as a relational view of the equal sign is an important component to this transition (e.g., McNeil & Alibali, 2006). If one style of instruction is superior at improving performance on other indicators of understanding, that instructional style has clear benefits. Across the three indicators of understanding that were examined, the pattern was consistent: Conceptual instruction led to greater performance than procedural instruction, and manipulatives did not have an effect. Therefore, although performance on equivalence problem solving was similar across conditions following instruction, conceptual instruction carried the advantage of increased generalization of learning.

In general, students improved in their ability to reconstruct problems following a brief presentation, and students in the conceptual conditions did particularly well. This improvement suggests that students began to encode problem features that they might have previously ignored (McNeil & Alibali, 2002). Encoding relevant problem features may indicate that a student is considering new, more appropriate strategies rather than simply applying the addition schema (McNeil & Alibali, 2004), and it is positively related to subsequent learning on equivalence problems.

Although only modest improvements were observed in the definition rating task, the improvements are notable given the difficulty researchers have had convincing children to relinquish their previously held belief about the operator meaning of the equal sign (e.g., Carpenter et al., 2003). Carpenter previously observed that explaining the meaning of the equal sign is not enough to change children's thinking. In this study the repeated lessons may have helped to convince students to modify their belief. It is impossible to determine the effect of the repeated instruction, however, because the indicators of understanding were not tested between Sessions 1 and 4 due to time constraints.

As previously discussed, a limitation of this study is the reliability of the measures used as indicators of understanding. A person's understanding of a concept is multi-dimensional and can be measured in many different ways. Even with perfect measures, each method of tapping "understanding" can point to different conclusions because it is tapping a different aspect or mode of demonstrating understanding (Bisanz et al., 2009). It is often impossible to simply conclude that a person "understands" or "does not understand," but the similar pattern across measures allow us to draw conclusions about the general effect of procedural versus conceptual instruction.

The lack of effect of manipulatives in any of the indicators of understanding or problem solving was somewhat surprising because in previous studies young children have been found to succeed on problems presented in blocks at younger ages than problems presented as Arabic numbers (e.g. Falkner et al., 1999; Sherman & Bisanz, 2009), and manipulatives have been found to be particularly successful at improving performance among Grade 2 students (Watchorn & Bisanz, 2005). Others, however, have found that blocks can be distracting for students (e.g., Uttal, Scudder, & DeLoache, 1997), so perhaps any benefits that some students gained from the use of the manipulatives was countered by a detrimental effect in others. Alternatively, the specific verbal and gestural instruction used in this study may have been elaborate enough to help students make the connections that the students in other studies were only able to make with the use of blocks. It is impossible to say with certainty why the manipulatives did not have an effect in this study, but this finding has important educational implications as many teachers go out of their way to include manipulatives with every math lesson.

Despite the overall superiority of the conceptual instructional focus and the lack of effect of manipulatives, it is difficult to determine whether procedural instruction or the inclusion of manipulatives might be key components to the success of certain individuals. Some of the students who did not improve in the conceptual conditions might have improved had they been given procedural instruction and vice versa. Similarly, the use of manipulatives may have helped some students but hindered others' performance, resulting in a lack of effect overall. It would be advantageous for educational purposes if teachers could identify which students would benefit from which method of instruction, but a different kind of design would be necessary to obtain this information.

The fourth question was whether any pre-existing skills are associated with readiness to learn how to solve equivalence problems. Indicators of readiness to learn can be informative for educators and for our understanding of cognitive development. A regression revealed that performance on the problem reconstruction and equation rating tasks in Session 1 was associated with post-instruction performance on the equivalence problem-solving task in Session 4. Thus, the results confirmed McNeil and Alibali's (2005b) finding that knowledge of which features of problems to encode is associated with readiness to learn, and further demonstrated that acceptance of non-canonical equations is associated to a limited extent with readiness to learn.

The fifth question was whether groups of children respond to equivalence-related tasks in qualitatively distinct ways. If we can identify characteristic profiles of performance, we might better be able to understand children's strengths and weaknesses and tailor instruction to their needs. Four clusters were found and, based on these clusters, a possible developmental progression was suggested. In this hypothesized progression children begin with an operational interpretation of the equal sign, then learn the relational definition of the equal sign, but only gradually learn to apply this new definition to understand non-canonical equations. An examination of the cluster composition based on instructional focus was consistent with the view that conceptual instruction may facilitate progression to a more advanced understanding of the equal sign, as the students who received conceptual instruction were the most likely to be in the cluster with the most advanced relational understanding of the equal sign. This proposed sequence is speculative, of course, and a longitudinal study would be required to test its validity.

The findings presented here have implications both for understanding cognitive development and for optimizing instruction. The debate about the most effective instructional approaches to enhance student performance in mathematics is of particular importance to those creating policy, designing, and selecting curricula for schools. At the fore of the debate is the question of whether student-centered approaches, such as CGI, or more teacher-directed approaches, such as the instruction provided in this study, are superior (Agodini, Harris, Thomas, Murphy, Gallagher, & Pendleton, 2010). The success observed here implies that more didactic instruction, which is likely easier for teachers to learn, may also be effective at improving student performance on equivalence problems. The concern remains that students' success may simply reflect rote learning, but rote learning may be an important first step in learning about mathematical equivalence.

The majority of interventions previously tested on equivalence problems involved only a short lesson, were administered in one session, and were implemented one-on-one rather than in a group setting. The instruction in this study more closely approximated classroom conditions by implementing the instruction in small groups and on two occasions. Granted, a certain degree of experimental control was necessary for this study, such as limiting the group sizes to smaller than is typical in classroom instruction. An important next step is to determine whether the instructional methods that were successful in this study would be equally successful outside of the controlled experimental environment. In future studies researches could partner with teachers to develop modules for practical implementation in the classroom based on the findings from this study.

If algebra is the "gatekeeper" to higher math, students must be well prepared to learn algebra so that future educational and employment opportunities are not constrained unnecessarily. The transition from arithmetic to algebra is one of the most significant hurdles students face in learning mathematics, but through an appropriate understanding of the principles that underlie both arithmetic and algebra, the transition need not be so dramatic and difficult. Equivalence is one such principle, and an appropriate understanding of the equal sign is crucial. By addressing this critical issue that has plagued Canadian and American children for so long, we can best prepare the next generation to face the challenges that lie ahead.

References

Agodini, R., Harris, B., Thomas, M. Murphy, R. & Gallagher, L. (2010). "Achievement Effects of Four Early Elementary School Math Curricula: Finding for First and Second Graders." Report submitted to the U.S. Department of Education, Institute of Education Sciences. Washington, DC: Mathematica Policy Research.

Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

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

Alibali, M. W., Knuth, E. J., Hattikudur, S., McNeil, N. M., & Stephens, A. C. (2007). A longitudinal examination of middle school students' understanding of the equal sign and equivalent equations. *Mathematical Thinking and Learning, 9*, 221-246.

Asquith, P., Stephens, A. C., Knuth, E. J., & Alibali, M. W. (2007). Middle school mathematics teachers' knowledge of students' understanding of core algebraic concepts: Equal sign and variable. *Mathematical Thinking and Learning, 9,* 247-270.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181–214.

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

Behr, M., Erlwanger, S., & Nichols, E. (1976). How children view equality sentences. PMDC Technical Report No. 3, Florida State University. (ERIC Document Reproduction Service No. ED144802).

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

Bisanz, J., & LeFevre, J. (1992). Understanding elementary mathematics. In J. I. D. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 113-136). Amsterdam, North-Holland: Elsevier Science Publishers.

Bisanz, J., Morrison, F. J., & Dunn, M. (1995). Effects of age and schooling on the acquisition of elementary quantitative skills. *Developmental Psychology, 31*, 221-236.

Bisanz, J., Sherman, J., Rasmussen, C., & Ho, E. (2005). Development of arithmetic skills and knowledge in preschool children. In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp.143-162). New York: Psychology Press.

Bisanz, J., Watchorn, R. P. D., Piatt, C., & Sherman, J. (2009). On "understanding" children's developing use of inversion. *Mathematical Thinking and Learning, 11,* 10-24.

Blume, G. W., & Heckman, D. S. (1997). What do students know about algebra and functions? In P. A. Kenney & E. A. Silver (Eds.), *Results from the* *sixth mathematics assessment* (pp. 225–277). Reston, VA: National Council of Teachers of Mathematics.

Bryant, P., & Nuñes, T. (2002). Children's understanding of mathematics. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 412-439). Malden, MA: Blackwell Publishers.

Byers, V., & Herscovics, N. (1977). Understanding school mathematics. *Mathematics Teaching, 81,* 24-27.

Capraro, R. M., Capraro, M. M., Ding, M., & Li, X. (2007). Thirty years of research: Interpretations of the equal sign in China and the USA. *Psychological Reports, 101,* 784-786.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. A. Romberg (Eds.) *Mathematics classrooms that promote understanding* (Chapter 2: pp. 19-32). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Carpenter, T.P., Levi, L., & Farnsworth, V. (2000). Building a foundation for learning algebra in elementary grades. *In Brief, 1.*

Carraher, D.W. & Schliemann, A. D. (2007). Early algebra and algebraic reasoning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of*

Teachers of Mathematics. Vol II (pp. 669-705). Charlotte, NC: Information Age Publishing.

Case, R. (1985). *Intellectual development: Birth to adulthood*. Orlando, FL: Academic Press.

Denmark, T., Barco, E., & Voran, J. (1976). Final report: A teaching experiment on equality. PMDC Technical Report No. 6, Florida State University. (ERIC Document Reproduction Service No. ED144805).

DeLoache, J. S. (2002). Early development of the understanding and use of symbolic artifacts. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 206-226). Malden, MA: Blackwell Publishing.

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

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27,* 6-11.

Flavell, J. H., & Wohlwill, J. F. (1969). Formal and functional aspects of cognitive development. In D. Elkind & J. H. Flavell (Eds.), *Studies in cognitive development* (pp. 67-120). New York: Oxford University Press.

Gallagher, A. M., De Lisi, R., Holst, P. C., McFillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology 75,* 165-190.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report, 22,* 1-22.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology, 91,* 720-730.

Goldin-Meadow, S., Wagner Cook, S., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math, *Psychological Science, 20,* 267-272.

Hattikudur, S. & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequalities help? *Journal of Experimental Child Psychology, 107,* 15-30.

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

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

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

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: Macmillan.

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.

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.

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.

McNeil, N. M. (2007). U-Shaped development in math: Seven-year-olds outperform nine-year-olds on equivalence problems. *Developmental Psychology, 43*, 687-695.

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

McNeil, N. M., & Alibali, M. W. (2000). Learning mathematics from procedural instruction: Externally imposed goals influence what is learned. *Journal of Educational Psychology, 92*, 734-744.

McNeil, N. M., & Alibali, M. W. (2002). A strong schema can interfere with learning: The case of children's typical addition schema. In C. D. Schunn & W. Gray (Eds.), *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society* (pp. 661-666). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

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.

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

Mix, K. S., (1999). Similarity and numerical equivalence: Appearances count. *Cognitive Development, 14*, 269-297.

Oksuz, C. (2007). Children's understanding of equality and the equal symbol. *International Journal for Mathematics Teaching and Learning,* 1-19.

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

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

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

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.

Rittle-Johnson, B., & Star, J. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99*, 561-574.

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

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at US mathematics and science education.* Dordrecht, the Netherlands: Kluwer.

Schneider, M., & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology, 46,* 178-192.

Seo, K., & 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: Lawrence Erlbaum Associates, Inc.

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

Siegler, R.S. (2006). Microgenetic analysis of learning. In W. Damon & R. M. Lerner (Series Eds.), D. Kuhn & R. Siegler (Vol. Eds.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (6th ed., pp. 464-510). New York: John Wiley & Sons.

Uttal, D. H., Scudder, K. V., & DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology, 18*, 37-54.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236.

Watchorn, R., & Bisanz, J. (2005, October). Enduring instructional effects on equivalence problems. Poster presented at the biennial meeting of the Cognitive Development Society, San Diego, CA.

Watchorn, R., Lai., M., & Bisanz, J. (2009, June). Canadian vs. Taiwanese Children's Performance on Equivalence Problems. Poster presented at the

Annual Convention of the Canadian Psychological Association, Montréal, Québec.

Wynroth, L. (1975). *Wynroth math program – the natural numbers sequence.* Ithica, N.Y.

Appendix A

Practice Problems

$4 + 5 + 3 = \_\_\_ + 6$

$6 + 4 + 3 = 6 + \_\_\_$

$4 + 3 + 6 = 3 + \_\_\_$

$3 + 4 + 5 = \_\_\_ + 3$

$5 + 3 + 7 = 5 + \_\_\_$

$6 + 4 + 5 = \_\_\_ + 6$

$4 + 5 + 6 = \_\_\_ + 4$

$3 + 7 + 5 = 7 + \_\_\_$

$1 + 3 + 2 = \_\_\_ + 2$

$9 + 1 + 1 = 9 + \_\_\_$

Appendix B

Sample Equivalence Test

Mathematics – Version 1
(Please circle the correct answer)

7 + 5 + 3 = ___ + 7			
6	22	8	15

6 + 7 = ___ + 5			
18	13	4	8

3 + 4 + 5 = 2 + ___			
14	12	8	10

5 + 6 + 4 = ___+ 4			
15	11	19	9

4 + 5 = 3 + ___			
6	4	9	12

6 + 4 + 3 = 5 + ___			
18	13	4	8

5 + 6 = ___ + 4			
3	15	7	11

5 + 3 + 7 = ___ + 4			
15	11	19	6

4 + 5 + 6 = ___ + 2			
15	13	17	9

5 + 3 + 4 = 5 + ___			
2	17	7	12

3 + 6 + 4 = 4 + ___			
9	5	13	17

7 + 8 = 6 + ___			
21	15	5	9

Appendix C

Definition Rating Task

$$5 + 3 + 3 = 2 + \underline{\qquad}$$

		Not so smart	Kind of smart	Very smart
A.	Ken said "=" means "both sides of the sign should have the same amount":	☹	😐	☺
B.	Wendy said "=" means "the answer goes next":	☹	😐	☺
C.	Peter said "=" means "all the numbers after it are small":	☹	😐	☺

Appendix D

Equation Rating Task

Equations

Please circle one face on each row to tell us if you think the equation makes sense:

		No - this doesn't make sense	Maybe yes, maybe no – I don't know	Yes – this makes sense
A.	5 + 1 = 4 + 2	☹	😐	☺
B.	1 + 3 = 4	☹	😐	☺
C.	2 + 2 = 4 – 1	☹	😐	☺
D.	7 = 7	☹	😐	☺
E.	9 = 9 + 6	☹	😐	☺
F.	5 + 4 = 9 + 6	☹	😐	☺
G.	7 = 5 + 2	☹	😐	☺
H.	4 = 5	☹	😐	☺
I.	2 + 5 = 9	☹	😐	☺
J.	3 + 3 = 9 - 3	☹	😐	☺

Appendix E

Scoring for the Equation Rating Task

For each problem type, children were categorized as Relational, Transitional, Operational, or Inconsistent based on the pattern of their responses across both the correct and incorrect problem of that type. It was impossible to determine in children guessed on any of the items, so the patterns were determined based on the logic a child would use to respond assuming she was not guessing.

For all "correct" equations, other than the Standard problems which are not included in the analysis, the likely explanation for children who judged a problem as "making sense" was that they had a relational view of the equal sign. If a child judged the equation as "not making sense," then the likely explanation for that decision is that he or she holds an operational view of the equal sign. If a child chose "I don't know," it indicates some flexibility in thinking about occasions in which the equal sign might be used, and we might consider this child's view of the equal sign to be "in transition."

For the "incorrect" equations, whenever possible the number following the equal sign corresponded to the answer one would obtain by adding up all of the numbers before the equal sign, a common strategy among children with an operator interpretation of the equal sign. This pattern was not possible for the One-term Identity problems. If a child judged an incorrect equation as "making sense," the likely explanation is that she holds an operator interpretation of the equal sign. If a child chose "I don't know", again it indicates some flexibility in thinking about occasions in which the equal sign might be used, and we might consider this child's view of the equal sign to be "in transition". For an incorrect problem that a child judges as "not making sense" it is more difficult to determine which view of the equal sign the child holds. Two possibilities exist: (a) she holds a relational view of the equal sign and correctly identifies that the two sides are not equal, or (b) she holds an operational view of the equal sign, but thinks the format is incorrect. Thus, it was necessary to examine the child's pattern of responses across both the correct and incorrect problems to determine which view she holds of the equal sign, or if she responded inconsistently.

A child was only categorized as holding a relational view of the equal sign for that problem type if he responded that the correct problem "made sense" and the incorrect problem "did not make sense". A child was categorized as "transitional" if he responded in way consistent with a transitional view on both problems of that type (i.e., selected "I don't know" for both), or in a way consistent with a relational view on one problem and transitional on the other. A child was categorized as "operational" if he responded in way consistent with an operational view on both problems of that type, or in a way consistent with a operational view on one problem and transitional on the other. Lastly, a child was categorized as "inconsistent" if he responded in a way consistent with an operational view on one problem and relational on the other. The inconsistent responses were coded as missing data as it is difficult to determine how the children were interpreting the task.