IS THAT CORRECT? CLARIFYING THE EFFECTS OF FEEDBACK DURING

MATHEMATICS PROBLEM SOLVING

By

Emily R. Fyfe

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

August, 2015

Nashville, Tennessee

Approved:

Professor Bethany Rittle-Johnson

Professor Lynn Fuchs

Professor Lisa Fazio

Professor Bruce McCandliss

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Feedback is a ubiquitous learning tool that has been studied by cognitive scientists,

learning theorists, and educational psychologists alike (e.g., Hattie & Timperley, 2007; Kluger &

DeNisi, 1996). It is broadly defined as any information about performance or understanding that

the learner can use to confirm, reject, or modify prior knowledge (Mory, 2004).

Feedback is often touted as an effective form of guidance (Alfieri, Brooks, Aldrich, &

Tenenbaum, 2011; Shute, 2008). However, the effects of feedback are quite variable. In some

cases it promotes learning, but in others it has neutral or negative effects relative to a no

feedback control (Kluger & DeNisi, 1996). Recent research suggests that learners' prior

knowledge may moderate the effects of feedback (e.g., Fyfe, Rittle-Johnson, & DeCaro, 2012).

Specifically, feedback may benefit learning and performance for those with low prior

knowledge, but hinder learning and performance for those with higher prior knowledge. The

results temper recommendations for feedback and suggest that while giving feedback can be

beneficial, withholding feedback may be more beneficial for certain learners.

Given the potential implications of these results for the role of feedback specifically and

for learning theories broadly, more research is needed to verify and clarify the conclusions. For

example, although several studies have identified learners' prior knowledge as a potential

moderator, no causal link between prior knowledge and feedback effects has been established.

Further, reasons underlying the negative effects of feedback remain unclear. Although several

mechanisms have been suggested, more empirical work is needed, particularly with learners who

already have some prior knowledge. In the present series of experiments, I investigated the roles of prior knowledge and feedback during mathematics problem solving for elementary-school children. Specifically, I examined the causal role of prior knowledge and also manipulated aspects of feedback to determine the conditions under which its impact was negative.

The Mixed Effects of Feedback

As mentioned previously, feedback is a broad term that refers to many types of information given to a learner. The amount of information can vary on a continuum from simple right-wrong verification to more elaborate explanations, such as a conceptual rationale of the correct answer or a hint at a correct problem-solving procedure. Typically, the goal of providing feedback is to reduce gaps between current and desired levels of knowledge. In the current series of studies, I focus on two types of feedback that are common on instructional tasks: right-wrong verification and right-wrong verification plus the correct answer.

There are practical, empirical, and theoretical reasons for the popularity of feedback. For example, feedback can be applied by parents and teachers in nearly any learning situation. Indeed, parents frequently provide feedback to their children on learning tasks at home (e.g., Evans, Barraball, & Eberlee, 1998; Hoover-Dempsey et al., 2001), and teachers also provide feedback on student performance in the classroom (e.g., Pianta, Belsky, Houts, & Morrison, 2007). Further, meta-analyses comparing feedback to no-feedback controls confirm its powerful influence on learning (e.g., Bangert-Drowns et al., 1991; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). For example, a recent analysis reported an average positive effect size of .46 for feedback relative to no feedback (Alfieri et al., 2011), indicating that feedback generally has beneficial effects on performance and learning in instructional settings. Feedback is theorized to

benefit learning in several ways. For example, it can reinforce or strengthen initially correct responses (Smith & Kimball, 2010). It can also help promote error-detection and correction. Specifically, feedback can reduce perseveration on initially incorrect responses (Kulhavy, 1977), and facilitate the generation of (potentially) correct alternatives (Butler & Winne, 1995). In general, feedback is often assumed to be helpful and many agree that "the importance of feedback in promoting learning is inarguable" (Moreno, 2004, p. 100).

Despite broad endorsement of feedback, research indicates that the effects of feedback vary considerably and are not universally beneficial (see Mory, 2004). For example, in two meta-analyses, feedback had positive effects in most cases, but negative effects in a third of the cases (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996). To be clear, negative effects do not mean that feedback reduces one's knowledge relative to an initial starting point. Rather, negative effects occur when feedback leads to lower learning outcomes compared to no feedback. Researchers have theorized that feedback has negative effects when it reduces mindfulness (e.g., relying solely on the feedback to provide the correct answers, Butler & Winne, 1995), draws attention to the self (e.g., evaluating one's abilities and self-image, Kluger & DeNisi, 1998), or produces cognitive interference (e.g., confusing one's response and the correct response, Kulhavy, 1977). However, the vast majority of past feedback research was with adults in laboratory contexts recalling test-like material (e.g., multiple-choice, list-learning). Further, learner characteristics that interact with feedback (i.e., moderators) have rarely been tested.

In response to the mixed impact of feedback, researchers have called for future work to specify key moderators of feedback with more diverse learning content and more diverse populations (Hattie & Gan, 2011). The current research focuses on the impact of feedback on children generating problem solutions. Previous findings on the effects of feedback in memory

tasks may not generalize to the effects of feedback in problem solving tasks. For example, generating multi-step solutions requires more complex cognitive processing relative to retrieving a stated fact (e.g., a word pair). Thus, integrating the feedback information may be more difficult. Further, elementary-school aged children often have lower working memory capacity, knowledge, and meta-cognition relative to adults, providing an additional reason why previous findings may not generalize. Thus, one goal of the current study was to extend previous research on feedback to a new task (problem solving) and a new population (elementary-school children). A second goal was to experimentally test one potential moderator to better predict when feedback will help versus harm learning within a task and population. Specifically, I tested the role of prior knowledge in learning from feedback and whether it moderated feedback effects.

The Role of Prior Knowledge

There are several reasons to suggest that learners' prior knowledge is a key moderator of the effects of feedback. First, nearly all theoretical models of feedback give the learner's prior knowledge a primary role (e.g., Mory, 2004; Narciss & Huth, 2004). Specifically, learning from feedback is often viewed as an interaction between information in long-term memory (i.e., prior knowledge) and the new information provided in the feedback message. Butler and Winne (1995) explicitly note that "feedback…is contextualized according to a student's prior knowledge" and that "a student's prior knowledge and beliefs inherently influence how learning proceeds" (p. 264). Indeed, many agree that "to be effective, feedback needs to be…compatible with students' prior knowledge" (Hattie & Timperley, 2007, p. 104).

Second, prior knowledge often moderates the effectiveness of various instructional techniques, resulting in classic aptitude-by-treatment interactions (Cronbach & Snow, 1977). For

example, a wide range of studies have demonstrated these interactions termed "expertise reversal effects," in which an instructional technique that is effective for low-knowledge learners loses its benefits for learners with higher prior knowledge (see Kalyuga, 2007). Expertise reversal effects have been found in a variety of domains, including mathematics problem solving (e.g., Kalyuga & Sweller, 2004). Importantly, a common conclusion stemming from this research is that "instructional guidance, which may be essential for novices, may have negative consequences for more experienced learners" (Kalyuga et al., 2003, p. 24). Thus, prior knowledge is a key feature to consider in the study and application of instructional guidance and feedback.

Third, evidence from multiple experimental studies suggests that prior knowledge often predicts learning from feedback (Fyfe et al., 2012; Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Krause, Stark, & Mandl, 2009; Luwel, Foustana, Papadatos, & Verschaffel, 2011; Nihalani, Mayrath, & Robinson, 2011). Specifically, feedback often does not help those with sufficiently high prior knowledge. For example, undergraduate students with low prior knowledge of statistics exhibited higher learning on a posttest if they received explicit correct-answer feedback during training than if they did not. However, those with higher prior knowledge performed just as well when feedback was not provided (Krause et al., 2009).

In problem-solving domains, prior knowledge of correct strategies seems particularly relevant. For example, several studies have targeted children who exhibit low knowledge of correct strategies at pretest (e.g., use incorrect strategies to solve pretest problems), and they find that feedback has positive effects on the generation of diverse strategies relative to no feedback (Alibali, 1999; Fyfe et al., 2012). In contrast, other studies have targeted adolescents with high knowledge of correct strategies at pretest, and they find that these students gain similar strategy knowledge whether feedback is provided or not (Hofer et al., 2011; Nussbaumer et al., 2014).

Together, these studies suggest that feedback has positive effects for learners with low knowledge, but neutral effects for those with higher prior knowledge of correct strategies.

More concerning is the possibility that feedback has *negative* effects for some learners. Preliminary evidence for this comes from two of our previous experiments in which second- and third-grade children solved novel math problems prior to receiving instruction (Fyfe et al., 2012). During problem solving, some children received feedback after each problem and others did not. For children with low prior knowledge of correct strategies, feedback facilitated posttest problem solving. But, for children with moderate prior knowledge, feedback hindered performance relative to no feedback. This occurred even though most of the "moderate-knowledge" children used correct strategies on less than 40% of problems at pretest. The effects were maintained two weeks later, and did not depend on feedback type. For example, feedback that provided the correct answer and feedback that provided right-wrong verification only yielded similar results. Similarly, feedback that focused on children's numerical answer (i.e., solution outcome) yielded similar effects as feedback that focused on how they obtained the answer (i.e., solution strategy).

Given the counterintuitive nature of these results, more work is needed to clarify their implications for theory and for practice. First, no causal link has been established between learners' prior knowledge and the effects of feedback on learning. Previous studies have relied on pre-existing indicators of prior knowledge (e.g., researcher-created pretests; Fyfe et al., 2012; Gielen et al., 2010; Krause et al., 2009). However, these learners may vary on a number of factors (e.g., motivation, intelligence) that influence their response to feedback. One solution is to use a pre-familiarization technique, in which some learners are exposed to the target material and others are not (Petersen & McNeil, 2013; Rey & Buchwald, 2011). This avoids confounding

variables, allows for random assignment to prior knowledge condition, and establishes the causal relation (not just association) between prior knowledge and feedback effects (Tobias, 2010).

A second issue is the need to specify the type and level of prior knowledge at which the reversal occurs. For example, at what level of knowledge is feedback detrimental? Past research has relied on a median split to classify low- and high-knowledge learners (e.g., Fyfe et al., 2012; Krause et al., 2009). Such a sample-specific, post-hoc approach does not allow for any pre-specified classification criteria. Yet, a priori predictions are necessary for good theory and for translation into practice. One might intuitively set the criteria as mastery of a correct strategy. Yet, our prior studies indicate that the negative effects of feedback occur for learners with only moderate knowledge of correct strategies (e.g., they used correct and incorrect strategies inconsistently), suggesting the threshold may be some versus no knowledge of a correct strategy.

A final issue is to better understand why feedback may have negative effects. Learners with moderate prior knowledge in the domain may be particularly susceptible to negative effects precisely because they can activate their knowledge during the task. Although generally helpful, knowledge-activation may have potential consequences as well. First, it may increase learners' expectation of performing well compared to learners with no prior knowledge, and thus heighten their sensitivity to feedback that states otherwise (Kluger & DeNisi, 1996). Second, greater knowledge-activation may increase the processing of redundant information. For example, for higher-knowledge learners, feedback provides information about a problem the learner already knows to a certain degree or can figure out with time. This redundancy may cause learners to spend cognitive resources on unnecessary information and reduce learning (Sweller, Ayres, & Kalyuga, 2011). In the current study, I included trial-by-trial microgenetic analyses and subjective student reports to better understand how feedback impacted the learning process.

Thus, one goal of the current study was to test the causal role of a pre-specified level of prior knowledge on the effects of feedback. In Experiment 1, I used a pre-familiarization technique to randomly assign learners to knowledge condition. Specifically, I provided some learners with knowledge of a correct strategy and other learners with no knowledge of a correct strategy. I predicted that children with no knowledge of a correct strategy would benefit from feedback, but that, contrary to conventional wisdom, children with induced knowledge of a correct strategy would be hindered by feedback relative to no feedback.

The Timing of Feedback

A second goal of the current study was to explore if the negative effects of feedback are robust across various feedback types – specifically feedback that varies in its timing. In previous experiments, the *immediacy* of the feedback may have heightened potential consequences of feedback for higher-knowledge learners. Immediate feedback is provided right after a learner has responded to a problem, and has the chance to impact (and potentially impede) ongoing processing of the current problem as well as performance on subsequent problems. For example, if immediate feedback provides redundant information and results in cognitive overload, learners may have fewer resources to process subsequent problems, thereby gaining little from the experience. One solution is to delay feedback rather than provide it on a trial-by-trial basis. For example, summative feedback is provided after the learner has responded to all problems in a set.

A large literature exists on the timing of feedback with substantial disagreement regarding the optimal timing. Researchers motivated by behavioral theories of reinforcement believe that feedback must be given as soon after a response as possible in order to eliminate incorrect ways of thinking and reinforce correct ones (e.g., Skinner, 1954). Further, immediate

feedback may provide motivation to practice, as progress can be continually monitored and goal-driven behavior can ensue (Shute, 2008). However, others believe delaying feedback is more beneficial. First, it may prevent learners from becoming over-reliant on the immediate presentation of the correct solution, which in turn may increase mindful processing and the need to exert effort on one's own response (Bangert-Drowns et al., 1991). Second, delaying feedback allows for the strength of initially incorrect responses to dissipate, which may make processing correct responses easier (Kulhavy, 1977). Finally, delaying feedback also allows for spaced presentation of information, as the problem or study material will be presented first during the initial response stage and later during the feedback presentation stage (Schmidt & Bjork, 1992).

Several meta-analyses point to the advantages of immediate feedback, particularly in computer based instruction (Azevedo & Bernard, 1995) and verbal learning on classroom quizzes (Kulik & Kulik, 1988). Indeed, multiple experimental studies have demonstrated the superiority of immediate feedback over delayed feedback for the acquisition of verbal materials and procedural/motor skills (Anderson et al., 2001; Corbett & Anderson, 1989; Dihoff, Brosvic, Epstein, & Cook, 2003). For example, Dihoff and colleagues (2003) had undergraduate students in a psychology course complete five multiple-choice quizzes during the semester. Using a within-subjects design, on each quiz students either received immediate, trial-by-trial feedback, summative feedback after the quiz, or summative feedback the next day. On a final exam at the end of the semester, students demonstrated the highest recall and lowest perseveration of incorrect responses on the test items on which they had originally received immediate feedback.

Despite the reported benefits of immediate feedback, other research has found advantages to delaying the presentation of feedback (Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Kulhavy, 1977; Kulhavy & Anderson, 1972). For example, Butler and

colleagues (2007) had undergraduate students study general knowledge passages and take a multiple-choice test. For some items, students received feedback immediately after selecting a response. For other items, the feedback was delayed until after the test was completed. Delayed feedback led to superior performance on a final cued-recall test, and the effect was stronger the longer the retention between feedback and testing. Indeed, experiments with motor tasks have consistently found that while immediate feedback results in more efficient learning, delayed feedback results in improved retention and fewer errors on subsequent assessments (Schmidt & Bjork, 1992). Often, proponents of delayed feedback acknowledge that immediate feedback offers several advantages, including clear guidance toward correct responses. However, the argument is that the disadvantages outweigh them (e.g., interrupt processing during acquisition phase, encourage over-reliance on the feedback message, massed presentation of information).

Thus, a second goal of the current study was to explore whether the negative effects of feedback for higher-knowledge learners were robust by manipulating the timing of feedback. Several researchers suggest that delaying feedback may be particularly beneficial for learners with higher knowledge in the target domain relative to lower knowledge learners (Mason & Bruning, 2001; Shute, 2008), though this has never been experimentally tested. In Experiments 2, 3, and 4, I used a pre-familiarization technique to ensure that all participants had knowledge of a correct strategy. Then I manipulated the presence and timing of feedback provided. I predicted that immediate feedback would have negative effects relative to no feedback (a replication of our previous work), and that summative feedback would have positive effects relative to no feedback. Although summative feedback may still produce cognitive or affective responses similar to immediate feedback, it is not provided *during* problem-solving when task-relevant

processing is ongoing. Further, as mentioned above, previous research has found benefits of delaying the presentation of feedback on subsequent assessments given after the initial learning.

The Current Study

The goal of the current study was to examine the roles of feedback and prior knowledge on mathematics problem solving. In Experiment 1, I examined the causal role of prior knowledge and whether differences in prior knowledge resulted in varying feedback effects. In Experiments 2, 3, and 4, I examined the presence and timing of feedback for learners with some prior knowledge. Across all experiments, I tested the impact of feedback in the context of elementary-school children learning to solve math equivalence problems (i.e., problems with operations on both sides of the equal sign, such as $3 + 4 + 5 = 3 +$ __).

There are several reasons I selected math equivalence as the target domain. First, it is an educationally-relevant and important topic for young children to learn. Math equivalence, typically represented by the equal sign, is the relation between two quantities that are equal and interchangeable (Kieran, 1981), and it is arguably one of the most important concepts for developing young children's algebraic thinking (Falkner, Levi, & Carpenter, 1999; Knuth, Stephens, McNeil, & Alibali, 2006). It is necessary for performing the same operation on both sides of an equation and for understanding equivalent expressions (Kieran, 1992; Steinberg, Sleeman, & Ktorza, 1990). Indeed, the Common Core State Standards recognize the importance of math equivalence and have included it in their standards as early as first grade. For example, one first grade standard (Standard 1.0A.7) is to "understand the meaning of the equal sign, and determine if equations involving addition and subtraction are true or false. For example, which of the following equations are true and which are false? $6 = 6$, $5 + 2 = 2 + 5$, $4 + 1 = 5 + 2$."

Second, math equivalence is a relatively difficult topic for elementary-school children, and they often have room to grow (Fyfe, DeCaro, & Rittle-Johnson, 2015; Fyfe, McNeil, & Borjas, 2015; McNeil, Fyfe, Dunwiddie, 2015; McNeil, Fyfe, Petersen, Dunwiddie, Brletic-Shipley, 2011; McNeil et al., 2012). Indeed, decades of research indicate that most U.S. children develop a poor understanding of math equivalence and fail to solve math equivalence problems correctly (e.g., McNeil, 2008; Rittle-Johnson & Alibali, 1999; Weaver, 1973). For example, across nine studies, McNeil (2005) found that the vast majority of children (about 80%) did not succeed on the problems. Many children solved the problems incorrectly by adding all the numbers in a problem or just the numbers before the equal sign (e.g., for $3 + 4 + 5 = 3 + \_\_$, answering 15 or 12 rather than 9). Poor performance often stems from misinterpretations of the equal sign as an operator symbol meaning "get the answer" or "the total," as opposed to a symbol relating two equal amounts (Behr et al., 1980; Kieran, 1981; McNeil & Alibali, 2005).

Third, solving math equivalence problems and related tasks are often novel for elementary school children. Elementary-school children receive little explicit instruction on the meaning of the equal sign and are often only exposed to limited problem structures. Indeed, the vast majority of problems are presented in an "operations = answer" structure, and math equivalence problems are not typically included in elementary mathematics curricula (e.g., McNeil et al., 2006; Powell, 2012). For example, a relatively recent analysis revealed that of all instances in a textbook series for Grades 1 – 6, equations with operations on both sides of the equal sign accounted for just 4% of all instances of the equal sign (Rittle-Johnson et al., 2011).

Finally, previous research suggests that feedback influences performance on math equivalence problems (e.g., Alibali, 1999; Fyfe et al., 2012). For example, my previous experiments on the role of feedback and prior knowledge were conducted with elementary

school students learning to solve math equivalence problems (Fyfe, DeCaro, & Rittle-Johnson, 2015; Fyfe et al., 2012). In the current study, I extended this previous research in an attempt to verify and clarify the stated conclusions. I was primarily interested in the effects of feedback and prior knowledge on children's problem-solving success at posttest, both on problems similar to those presented during the intervention (i.e., learning items) and on problems with novel features (i.e., near transfer items). I also measured children's conceptual knowledge at posttest (e.g., understanding the meaning of the equal sign), although I expected feedback to have less of an impact on conceptual knowledge based on results from prior work (Fyfe et al., 2012).

CHAPTER II

EXPERIMENT 1

The goal of Experiment 1 was to test the causal role of prior knowledge. Although previous research suggests that the effects of feedback may depend on learners' prior knowledge, these studies have relied on pre-existing indicators of prior knowledge. Thus, it remains unclear whether it is prior domain knowledge per se that alters the impact of feedback or some other factor (e.g., motivation, intelligence, etc.) that is associated with prior knowledge.

In the study, I manipulated children's strategy knowledge prior to problem solving by providing some children with instruction on a correct problem-solving strategy and other children with instruction on a filler task. I also manipulated the provision of right/wrong verification feedback during problem solving by providing some children with verification feedback after each problem and other children with no feedback throughout the problem-solving task. I provided simple verification feedback (i.e., right/wrong judgment) for several reasons. First, I was interested in studying the effects of seemingly minor input during problem solving. In particular, given concerns that extensive feedback might overwhelm or disrupt ongoing cognitive processing, I opted to minimize the amount of information in the feedback message. Second, previous research has found substantial benefits of right/wrong verification feedback on children's problem solving (e.g., Bohlmann & Fenson, 2005; Brainerd, 1972). Third, in prior work, I have found that the content of feedback did not matter. For example, both verification feedback and verification plus the correct answer had positive effects for low-knowledge learners and negative effects for moderate-knowledge learners relative to no feedback (Fyfe et al., 2012).

14

I predicted that children with no initial knowledge of a correct strategy would benefit from immediate, verification feedback relative to no feedback. However, contrary to conventional wisdom, I predicted that children with induced knowledge of a correct strategy would be hindered by immediate, verification feedback relative to no feedback. This experiment is currently published in Fyfe and Rittle-Johnson (*in press*).

Method

Participants

Initial participants were 159 children from second- and third-grade classrooms in two public schools and one private school. The first public school served ethnically-diverse students in kindergarten through fourth grade (59% White, 24% Black, 10% Hispanic, 7% Asian), around half of whom (52%) received free or reduced priced lunch. In 2014, 45% of students in third- and fourth-grade scored proficient or higher on the state's standardized math assessment (Tennessee Comprehensive Assessment Program, TCAP). The second public school served primarily ethnic minority students in prekindergarten through fourth grade (13% White, 78% Black, 7% Hispanic, 2% Asian), most of whom (90%) received free or reduced priced lunch. In 2014, 33% of students in third- and fourth-grade scored proficient or higher on the TCAP math assessment. The small, private school primarily served middle- to upper-class White students in kindergarten through eighth grade with one class per grade.

Of the initial 159 children with parent consent and student assent, 112 (70%) met criteria for participation because they could not solve any math equivalence problems correctly (out of 4) on a screening measure. This ensured that any effects due to knowledge level were a result of the

15

strategy knowledge manipulation and not preexisting differences. Data from four additional children were excluded for failing to complete all activities (one child asked to stop halfway through the intervention session, two children ran out of time due to slow counting and weak arithmetic skills, and one student ran out of time because her session was interrupted by a school security drill). The final sample contained 108 children (*M* age = 8.4 yrs, *min* = 7.2 yrs, *max* = 9.8 yrs; 67 girls, 41 boys; 51 third-graders, 57 second-graders).

Design

The study had a 2 (induced strategy knowledge: yes vs. no) x 2 (feedback: present vs. absent) between-subjects design with children randomly assigned to conditions: (1) strategy knowledge with feedback (*n* = 27), (2) strategy knowledge without feedback (*n* = 26), (3) no strategy knowledge with feedback (*n* = 27), and (4) no strategy knowledge without feedback (*n* = 28). There were no differences between conditions in terms of age, gender, or grade (*p*s > .5).

Materials

Screening Measure. The screening measure was three tasks that tap understanding of math equivalence (from McNeil, Fyfe, Petersen, Dunwiddie, & Brletic-Shipley, 2011). See Appendix A for a list of the items and scoring criteria. For *equation solving*, children solved four math equivalence problems. Inclusion criterion was based solely on equation solving, as I was interested in children's knowledge of solution strategies. The two remaining tasks allowed me to test if conditions were matched on different aspects of prior knowledge. For *equation encoding*, children reconstructed four math equivalence problems after viewing each for five seconds to assess how they represented the structure of the problem (McNeil & Alibali, 2005). For *defining the equal sign*, children provided a written definition of the equal sign to assess their explicit, relational understanding (e.g., "same amount") of the symbolic representation for equivalence.

Intervention Problems. The 12 intervention problems (from Fyfe et al. 2012) included nine math equivalence problems with operations on both sides of the equal sign, with the unknown after the equal sign (e.g., $3 + 7 = \_\_ + 6$) or at the end (e.g., $5 + 3 + 9 = 5 + \_\_$). Three easier problems had an operation on the right side only (e.g., $9 = 6 + \_\_$).

Posttest. The posttest, adapted from past work (Matthews et al., 2012; Rittle-Johnson et al., 2011), was a broader measure that assessed children's procedural and conceptual knowledge of mathematical equivalence. See Appendix B for a list of the items and scoring criteria. The procedural knowledge scale included 8 items that assessed children's use of correct strategies to solve math equivalence problems ($\alpha = .90$). Half of the items were similar to those presented during the intervention (i.e., learning items) and half differed on a key problem feature, such as inclusion of subtraction (i.e., near transfer items). The conceptual knowledge scale included 10 items that assessed two key concepts: the relational meaning of the equal sign and the structure of equations ($\alpha = .73$). Given the central focus of the research on problem solving and strategy use, the secondary conceptual knowledge results are reported in full in Appendix C. Feedback condition did not significantly impact conceptual knowledge at posttest or retention test.

Coding. On the screening measure and posttest, children's problem-solving strategies were coded from their numerical answers. As in prior work (McNeil, 2008), responses within +/- 1 of the correct answer were coded as reflecting a correct strategy. For example, for the problem $2 + 7 = 6 + \_\_$, an answer of 15 indicated an incorrect "add all" strategy and an answer of 3 indicated a correct strategy. On the intervention problems, strategies were based on children's verbal reports (see Table 1 for example strategy reports). A second rater coded 30% of the responses and inter-rater agreement on specific strategy use was high (kappa = .94). Although we coded specific strategy use on the assessments (see Table 1 for specific strategy types), scores

were based solely on whether the strategy was correct or incorrect. Inter-rater agreement on this more basic coding was near perfect (kappa = .99).

Cognitive Load. Children's cognitive load was assessed using a 3-item task difficulty measure I have adapted from previous measures (Hart & Staveland, 1988; Paas et al., 2003) to be suitable for young children. I initially administered a 9-item measure with three subscales: mental effort, mental frustration, and task difficulty. Children responded to each item by circling their answer on a 4-point scale ranging from *strongly disagree* to *strongly agree*, with three items per subscale. Because it was a new measure for use with children, I explored the reliability and validity of the scale, including its relation to an existing single item used to measure cognitive load in adults (see Appendix D). The overall scale and some subscales resulted in items with low item-total correlations. However, the task difficulty subscale was both reliable and valid, so children's scores on the task difficulty scale were used to analyze subjective cognitive load.

Table 1. Strategies used to solve math equivalence problems

Strategy	For the problem $4 + 5 + 3 = 4 + \_\_$	
	Solution	Example Verbal Report
Correct Strategies		
Equalize	8	"4 plus 5 plus 3 is 12 and 4 plus 8 is 12."
Add-Subtract	8	"I added 4 plus 5 plus 3 and took away 4 from that."
Grouping	8	"I saw the 4 and the 4 and I just added the 5 and the 3."
Incorrect Strategies		
Add-All	16	"I just added them all up."
Add-to-Equal	12	"I added 4 plus 5 plus 3 and that equals 12."
Add-Two	9	"I added 4 plus 5 and that is 9."
Carry	5	"I saw 4 plus 5 here so I made it 4 plus 5 over here."

Procedure

Children completed the screening measure in their classrooms in a 10-minute session. Those who met the inclusion criteria then participated in a one-on-one tutoring intervention in a

single session lasting approximately 50 minutes. This session was conducted in a quiet area at the school with the author. The one-on-one session included four components: knowledge manipulation, knowledge check, problem solving, and an immediate posttest. See Appendix E for a complete script of the one-on-one intervention session.

Knowledge Manipulation. Children assigned to the *strategy-knowledge* condition received instruction on a correct strategy with four math equivalence problems presented on a computer one at a time. Two problems had the blank immediately following the equal sign (e.g., $5 + 4 + 3 = \_\_ + 5$) and two had the blank at the end (e.g., $3 + 4 + 2 = 3 + \_\_$). I used two problem types to increase the generalizability of the strategy (Matthews & Rittle-Johnson, 2009). Children were instructed on the equalize strategy, which involves adding the numbers on one side of the equal sign and then counting up from the number on the other side to get the same amount. I taught the equalize strategy as it is the strategy children tend to generate first on their own (e.g., Alibali, Phillips & Fischer, 2009). The experimenter provided instruction and demonstrated the procedure on all four problems. Children were asked to answer questions (e.g., "if we add up this side, what do we get?") to ensure they were attending to instruction.

Children assigned to the *no-knowledge* condition received instruction on a filler task (adapted from Hattikudur & Alibali, 2010) to control for time on task, interaction with the experimenter, and practice with addition. Children were directed to look at two boxes on the computer screen. One contained a single digit (e.g., 9) and the other contained a pair of addends (e.g., $3 + 4$). The pairs of addends were selected to match the computations required for the problems in the strategy-knowledge condition. Children were taught to decide which box has the biggest total number. The experimenter demonstrated the procedure on all four problems.

19

Children were asked to add the pair of addends for each problem and answer questions (e.g., "what is the next step?") to ensure they were attending to the task.

Knowledge Check. To ensure the knowledge manipulation worked, all children then solved a math equivalence problem on their own (i.e., $7 + 6 + 2 = 7 +$ __) and reported how they solved the problem. In the *no-knowledge* condition, children did not receive any feedback and simply moved on to the next task. In the *strategy-knowledge* condition, children were told whether or not they solved the problem using a correct strategy. If they used an incorrect strategy, instruction on the equalize strategy was repeated and they were asked to solve another problem until they solved one using a correct strategy and received feedback that it was correct. This procedure ensured that children in the strategy-knowledge condition had knowledge of a correct strategy and were aware that they used it correctly. I set the protocol such that after five failed attempts the experiment was discontinued and children received more remedial tutoring.

Problem Solving. Children were then asked to solve 12 math problems presented one at a time on a computer. See *Materials* for a description of the problems. After each problem, children reported how they solved the problem and either did or did not receive feedback.

In the *no-feedback* condition, children did not receive feedback and were simply told to go to the next problem. In the *feedback* condition, children received right/wrong verification feedback on their answer to the problem (e.g., "Good job! You got the right answer." / "Good try, but you did not get the right answer."). The feedback was based solely on the child's numerical answer and did not depend on the strategy reported. On a few occasions, children reported using a correct strategy, but obtained an incorrect answer due to an arithmetic mistake and received negative feedback. However, mismatches were rare (6% of all trials) and exclusion of children who experienced more than two mismatches ($n = 9$) did not impact the results.

Following the problem solving, children completed the cognitive load measure and the posttest. Children also completed a delayed retention test, which was identical to the posttest. There was large variability in the delay due to absences and school breaks, with a minimum delay of 6 days and a maximum delay of 34 days ($M = 15.2$ days, $SD = 4.5$). Indeed, only 58% of the sample completed the retention test in the intended two-week delay (± 2 days). Further, three children never completed it. Results are reported below for all children who completed a retention test, but they should be interpreted with caution given issues with administration.

Data Analysis

To examine children's performance on the primary outcome measures, I performed a series of ANCOVAs with strategy knowledge (yes vs. no) and feedback (present vs. absent) as between-subject variables. I included children's age and their scores on the screening measure as covariates. Preliminary analyses revealed no interactions with age or screening measure scores so these terms were not retained in the final models. I also explored the impact of grade on learning outcomes by including it as an additional between-subjects variable. There were general trends indicating that third-grade children often scored somewhat higher than second-grade children, but these main effects of grade rarely reached statistical significance and grade never interacted with feedback. Thus, grade was not retained in the final models.

Results

Screening Measure

Because of the inclusion criteria, all children in the final sample solved zero problems correctly on the equation-solving task on the screening measure. Some children succeeded on the

other two tasks. On average, children in the final sample encoded 1.6 (SD = 1.3) math equivalence problems correctly (out of 4), and 11% of children provided a relational definition of the equal sign. Importantly, performance on the equation encoding task and on the equal sign definition task did not differ as a function of condition, ps > .3. In line with previous research (Chesney et al., 2014; McNeil et al., 2011), I created a composite measure of children's performance. Specifically, I summed z-scores across the two task (Chesney et al., 2014). Composite scores ranged from −1.64 to 4.69 (M = 0.00, SD = 1.39) and did not differ as a function of condition, p = .97. Composite scores served as a covariate in subsequent analyses, though the pattern of findings was the same if it was not included.

Knowledge Check

The knowledge manipulation was successful. All children in the strategy-knowledge condition exhibited knowledge of a correct strategy after the manipulation; 88% of them used a correct strategy to solve the first problem, 6% needed a second problem, and 6% needed a third attempt. Nine children (16%) in the no-knowledge group solved the problem correctly, despite receiving no instruction on a correct strategy. Because these children could not be considered to have no knowledge of a correct strategy, they were excluded from analyses, leaving the no-knowledge group with 24 children in the feedback condition and 22 in the no-feedback condition. The pattern of findings was the same if these nine children were included.

Intervention Measures

For analyses of children's verbal strategy reports, I focused on the 9 math equivalence problems, which have operations on both sides of the equal sign. These problems elicit more easily identified strategies than problems with operations on only one side of the equal sign. However, the pattern of results was similar if I considered all 12 intervention problems.

Correct Strategy Use. I examined the percentage of trials on which children reported using a correct strategy. See Table 1 for example strategy reports. There was no main effect of feedback, $p = .45$. There was a main effect of knowledge, $F(1, 93) = 156.24, p < .001, \eta_p^2 = .63$, which was qualified by a significant feedback by knowledge interaction, $F(1, 93) = 11.13, p = .001, \eta_p^2 = .11$. For the strategy-knowledge group, children who received feedback used a correct strategy *less often* ($M = 77\%, SE = 5\%$) than children who did not ($M = 90\%, SE = 5\%$), $F(1, 93) = 3.51, p = .06, \eta_p^2 = .04$, though this difference did not reach significance. For the no-knowledge group, children who received feedback used a correct strategy significantly *more often* ($M = 30\%, SE = 5\%$) than children who did not ($M = 9\%, SE = 5\%$), $F(1, 93) = 7.87, p = .01, \eta_p^2 = .08$.

A trial-by-trial examination of children's correct strategy use supports these conclusions (see Figure 1). For the strategy-knowledge group, children who did not receive feedback performed consistently well across the items. In contrast, children who received feedback did well on easier items (e.g., 3 + 7 = 3 + __), but their performance suffered on the next, more difficult item (e.g., 3 + 7 = __ + 6). For the no-knowledge group, the pattern was the opposite. Feedback seemed to help no-knowledge children use an easier item to help with the next difficult item. In contrast, those who did not receive feedback did poorly on the difficult items.

Correct Strategy Generation. I also examined the types of strategies children reported using (see Table 1). For each child, I calculated the number of different types of correct strategies he/she used. As long as children reported the strategy on at least one problem, they received credit for using it. The number of different correct strategies used during the intervention ranged from zero to three ($M = 1.1, SD = 0.8$). There was no effect of feedback, $p = .17$. There was an effect of knowledge, $F(1, 93) = 23.65, p < .001, \eta_p^2 = .20$, which was qualified by a significant feedback by knowledge interaction, $F(1, 93) = 6.30, p = .01, \eta_p^2 = .06$. For the

strategy-knowledge group, children generated a similar number of correct strategies whether they received feedback ($M = 1.3$, $SE = 0.1$) or not ($M = 1.4$, $SE = 0.1$), $p = .42$. For the no-knowledge group, children who received feedback generated a greater number of correct strategies ($M = 1.0$, $SE = 0.1$) than children who did not ($M = 0.4$, $SE = 0.1$), $F(1, 93) = 7.13$, $p = .01$, $\eta_p^2 = .07$.
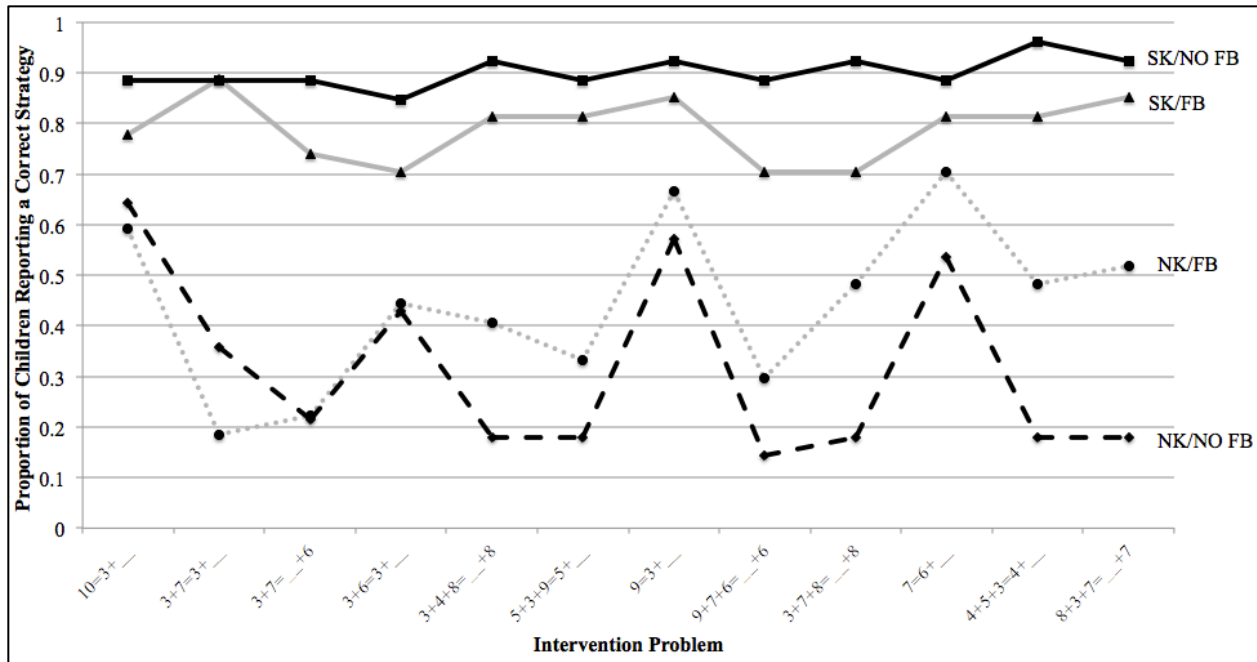
Note. SK: Strategy Knowledge condition. NK: No Knowledge condition. FB: Feedback condition. NO FB: No Feedback condition. Intervention problems are listed in the order in which children solved them during the problem-solving task.

Figure 1. Correct strategy use on each intervention item in Experiment 1

Incorrect Strategy Generation. For each child, I also calculated the number of different types of incorrect strategies he/she used. The number of different incorrect strategies ranged from zero to four ($M = 1.0$, $SD = 0.9$, see Table 1). There was no feedback by knowledge interaction, $p = .76$. However, there were main effects of feedback, $F(1, 93) = 4.43$, $p = .04$, $\eta_p^2 = .05$, and knowledge, $F(1, 93) = 64.24$, $p < .001$, $\eta_p^2 = .41$. Children who received feedback generated a greater number of incorrect strategies ($M = 1.2$, $SE = 0.1$) than children who did not

($M = 0.9$, $SE = 0.1$). Children in the no-knowledge group generated a greater number of incorrect strategies ($M = 1.6$, $SE = 0.1$) than children in the strategy-knowledge group ($M = 0.4$, $SE = 0.1$). There were also differences in perseveration—reporting the same incorrect strategy on all nine math equivalence problems. Nine children (41%) in the no-knowledge without feedback condition perseverated, whereas none of the children in the three remaining conditions did.

Positive vs. Negative Feedback. The majority of children in the feedback conditions received a mix of both positive (i.e., that's correct) and negative (i.e., that's incorrect) feedback. For the strategy-knowledge group, a small portion ($n = 6$) received positive feedback on all twelve trials, but no children received negative feedback on all twelve trials. Thus, the negative effects of feedback were not due to children consistently being told they were wrong. Conversely, in the no-knowledge group, a small portion ($n = 5$) received negative feedback on all twelve trials, but only one child received positive feedback on all twelve trials. Thus, the positive effects of feedback were not due to children consistently being told they were right.

Cognitive Load. I also examined cognitive load by analyzing children's subjective reports of task difficulty. There were no main effects of feedback, $p = .23$, or knowledge group, $p = .35$. There was a significant feedback by knowledge interaction, $F(1, 93) = 4.81$, $p = .03$, $\eta_p^2 = .05$. For the strategy-knowledge group, children reported similar levels of cognitive load whether they received feedback ($M = 2.1$ out of 4, $SE = 0.2$) or not ($M = 2.3$, $SE = 0.2$), $p = .48$. For the no-knowledge group, children who received feedback reported higher levels of cognitive load ($M = 2.7$, $SE = 0.2$) than children who did not ($M = 2.1$, $SE = 0.2$), $F(1, 93) = 5.46$, $p = .02$, $\eta_p^2 = .06$.

Intervention Summary. During the intervention, right-wrong verification feedback had positive effects for low-knowledge children. Specifically, for children in the no-knowledge condition, feedback increased the frequency of correct strategy use, prevented perseveration on

25

the same incorrect strategy, facilitated the generation of more diverse strategies (both correct and incorrect), and also led to increased ratings of task difficulty relative to no feedback. In contrast, right/wrong verification feedback had neutral or negative effects for higher-knowledge children. Specifically, for children in the strategy-knowledge condition, feedback had no significant effect on the frequency of correct strategy use, it facilitated the generation of more incorrect strategies, and it led to similar ratings of task difficulty relative to no feedback.

Posttest Procedural Knowledge

Overall, children's performance on the posttest varied. Scores on the procedural knowledge scale ranged from 0 to 100% correct with approximately half of the children scoring higher than 50% and the other half scoring below 50%. Figure 2 displays children's procedural knowledge at posttest by condition. There was no main effect of feedback, $p = .30$. There was a main effect of knowledge, $F(1, 93) = 36.76$, $p < .001$, $\eta_p^2 = .28$, which was qualified by a large, significant feedback by knowledge interaction, $F(1, 93) = 16.80$, $p < .001$, $\eta_p^2 = .15$.

To interpret the interaction, I examined the effects of feedback for each knowledge group. For the strategy-knowledge group, children who received feedback exhibited significantly *lower* procedural knowledge ($M = 62\%$, $SE = 6\%$) than children who did not ($M = 81\%$, $SE = 6\%$), $F(1, 93) = 4.96$, $p = .03$, $\eta_p^2 = .05$. For the no-knowledge group, children who received feedback exhibited significantly *higher* procedural knowledge ($M = 50\%$, $SE = 6\%$) than children who did not ($M = 18\%$, $SE = 7\%$), $F(1, 93) = 12.38$, $p < .001$, $\eta_p^2 = .12$. Indeed, the effects of feedback were so positive for the no-knowledge group and so negative for the strategy-knowledge group that there were no statistical differences in the feedback condition between no-knowledge children ($M = 50\%$, $SE = 6\%$) and strategy-knowledge children ($M = 62\%$, $SE = 6\%$), $p = .16$. Exploratory analyses revealed that the results remain unchanged after excluding children

who received only positive feedback or children who received only negative feedback. Further, the pattern of results was consistent across the learning and transfer items. Overall, children with no knowledge of a correct strategy benefited from verification feedback relative to no feedback, but, for children with induced knowledge of a correct strategy, the reverse was true.
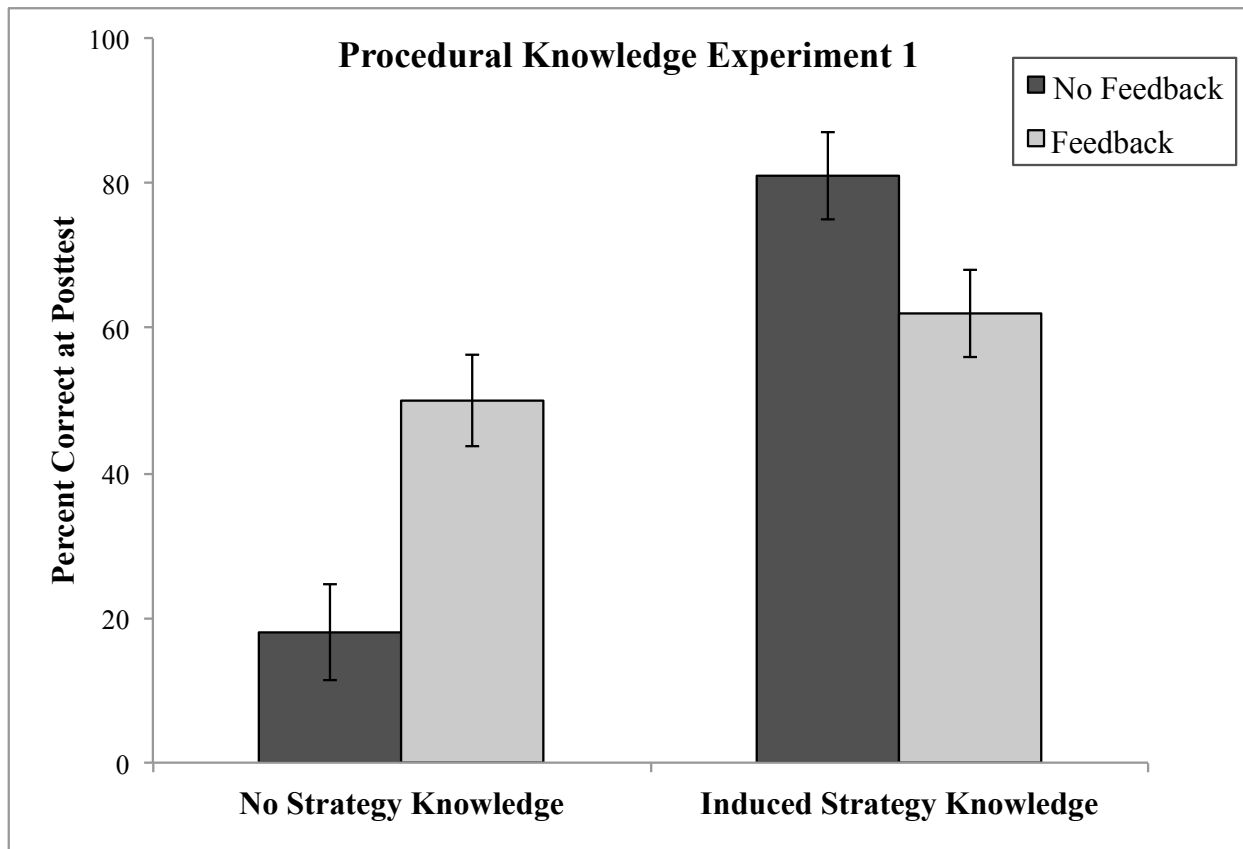
Note. Scores are estimated marginal means. Error bars represent standard errors.

Figure 2. Procedural knowledge at posttest in Experiment 1

Retention Test Procedural Knowledge

The pattern of results was somewhat similar at the retention test, though no longer reliable. There was no effect of feedback, $p = .46$. There was a main effect of knowledge, $F(1, 90) = 6.97$, $p = .01$, $\eta_p^2 = .07$, with children in the strategy knowledge condition exhibiting higher

procedural knowledge (M = 61%, SE = 5%) than children in the no-knowledge condition (M = 42%, SE = 6%). However, there was no reliable feedback by knowledge interaction, p = .31. Children in the strategy-knowledge condition exhibited similar procedural knowledge at the retention test whether they received feedback (M = 60%, SE = 7%) or not (M = 63%, SE = 8%). Children in the no-knowledge condition exhibited somewhat higher procedural knowledge when they received feedback (M = 47%, SE = 7%) than when they did not (M = 36% SE = 8%).

Discussion

Experiment 1 is the first study to provide causal evidence for the moderating role of prior knowledge on the effects of feedback. During mathematics problem solving, children with no knowledge of a correct problem-solving strategy benefited from right/wrong verification feedback relative to no feedback. In contrast, children who were taught a correct strategy learned more if they did not receive verification feedback. This effect occurred on children's problem-solving performance during the intervention and at posttest. The results confirm that feedback can have negative effects for children with some prior domain knowledge (e.g., Fyfe et al., 2012). Further, this negative effect can occur well before children exhibit mastery in the domain.

The results also provide insight into the mechanisms by which feedback helps low-knowledge learners; it prevented them from perseverating on the same incorrect strategy and helped them generate at least one correct one. Changes in strategy use may also explain the negative effects of feedback for those who had been taught a correct strategy. Specifically, feedback facilitated the generation of incorrect strategies, regardless of strategy knowledge. For

the strategy-knowledge children, this negative effect was not offset by any positive functions, thus, the introduction of incorrect strategies into their repertoires likely had detrimental effects.

The negative effects of feedback dissipated by the two-week retention test. One possibility is that feedback has neutral rather than negative effects for higher-knowledge learners in the long run. However, due to scheduling issues, only 58% of the sample completed the retention test so these results should be interpreted with caution given issues with administration.

Despite the positive contributions, several limitations remain. For example, it remains unclear whether the negative effects of feedback are specific to minimal verification feedback or whether they generalize to more informative feedback types. Indeed, comprehensive reviews indicate that feedback that provides verification and the correct answer is often more beneficial than verification alone (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996). Further, the immediacy of the feedback may have triggered cognitive or affective responses that interfered with task-relevant processing (e.g., Butler & Winne, 1995; Kluger & DeNisi, 1996). One possibility is that delaying feedback may reduce these negative, interfering effects.

To address these concerns, a second experiment was conducted similar to Experiment 1, but with several key modifications. First, given my interest in understanding the negative effects of feedback, I provided strategy instruction to all participating children to induce some strategy knowledge in all learners. Second, I employed correct-answer feedback (e.g., right/wrong judgment plus the correct answer) rather than simple right/wrong verification feedback to ensure the negative effects of feedback were not specific to verification feedback alone. Third, I manipulated both the presence and timing of feedback by including a no feedback condition, an immediate feedback condition, and a summative feedback condition.

CHAPTER III

EXPERIMENT 2

The goal of Experiment 2 was to provide additional insight into the negative effects of feedback for learners with some prior knowledge. In the study, I focused exclusively on learners with some prior knowledge by using a pre-familiarization technique. Specifically, prior to problem solving, I provided all children with instruction on a correct problem-solving procedure to ensure they had knowledge of a correct strategy. Then, during the subsequent problem-solving task, I manipulated the presence and timing of feedback. Specifically, children received no feedback, trial-by-trial immediate feedback after each problem, or summative feedback after the whole problem set. The feedback contained both verification (right/wrong judgment) and the correct answer to ensure that the negative effects of feedback in Experiment 1 were not due to the specific type of feedback used (i.e., minimal, verification-only feedback).

In line with Experiment 1, I predicted that immediate correct-answer feedback would have negative effects relative to no feedback for these higher knowledge children. However, I predicted that summative correct-answer feedback would have positive effects relative to no feedback. Delaying feedback until after problem solving should reduce any interfering effects that prevent the learner from attending to and learning from the task at hand. Further, it provides learners with valuable information regarding their problem-solving performance that they may be able to use on subsequent problem-solving tasks and assessments of knowledge. This experiment is currently published in Fyfe and Rittle-Johnson (*in press*).

Participants

Initial participants were 131 second- and third-grade children, none of whom participated in Experiment 1. Children were recruited from the two public schools in Experiment 1 as well as an additional public school. The third public school served a predominantly middle-class White population (80% White, 14% Black, 2% Hispanic, 3% Asian), with a small minority (18%) receiving free or reduced priced lunch. In 2014, 82% of students in third- and fourth-grade scored proficient or higher on the TCAP math assessment.

Of the initial 131 children, 113 (86%) met criteria for participation because they scored below 80% on a problem-solving screening measure. This criteria was adopted from my previous feedback study (Fyfe et al., 2012) and ensured that children had room to learn from the intervention. I used a more lenient inclusion criteria relative to Experiment 1 because all children in this study were given instruction on a correct strategy so it was not necessary that they started at the same initial knowledge level. Data from 12 additional children were excluded for failing to complete all activities (three students moved away, two student asked to stop halfway through the intervention session, five students ran out of time due to slow counting and weak arithmetic skills, and two students were special needs cases whom the teachers decided should not participate). The final sample contained 101 children (*M* age = 8.2 yrs, *min* = 7.0 yrs, *max* = 9.8 yrs; 57 girls, 44 boys; 72 third-graders, 29 second-graders).

Design

The study had a between-subjects design with children randomly assigned to conditions: (1) no feedback ($n = 33$), (2) immediate feedback ($n = 35$), and (3) summative feedback ($n = 33$). There were no differences between conditions in terms of age, gender, or grade, ps > .45.

Materials

The materials were identical to those in Experiment 1 with three exceptions. First, the screening measure included an additional, simpler equation-solving problem (i.e., $7 = \_\_ + 3$) to increase the variability in equation-solving scores. Second, to assess children's cognitive load, I only administered the validated three-item task difficulty scale. Third, to explore a potential reason for the negative effects of feedback, I also measured children's self-assessment (i.e., whether they considered their performance to reflect negatively on their traits and abilities) using a four-item measure used with kindergarten students in Kamins and Dweck (1999). Children were asked whether the task made them feel like they were good or not good at solving the problems, a good or a not good student, a nice or a not nice student, and a smart or a not smart student. Children received one point each time they chose the positive attribute and scores were summed to form an index ranging from 0 to 4. Internal consistency was sufficient ($\alpha = .78$), but the range of responses was restricted, as 82% of children always selected the positive attribute.

Procedure

The procedure was identical to Experiment 1 with a few exceptions. First, all children received instruction on a correct problem-solving strategy. Second, for the knowledge check problem, children were not told whether they solved the problem correctly. This ensured that children in the no-feedback condition never received feedback. If they solved the problem correctly, they were simply told to move on to the next activity. If they solved it incorrectly,

general instruction on the equalize strategy was repeated without revealing the correct answer, and they were asked to solve another problem until they solved one correctly (although the children were not told this criteria). Third, there was no retention test in this experiment given the data collection issues and weak effects on the retention test in Experiment 1.

Fourth, I manipulated both the presence and timing of feedback. The *no-feedback* condition was identical to Experiment 1. In the *immediate-feedback* condition, children received trial-by-trial correct-answer feedback, which included verification (as in Experiment 1) and also the correct answer. In the *summative-feedback* condition, children received verification and correct-answer feedback after all 12 problems had been solved. The problems with the child's solutions re-appeared on the computer screen, and the experimenter provided correct-answer feedback for each problem. The content and means of providing feedback were the same in the two feedback conditions with one exception. In the summative-feedback condition, the problems (with the child's solution) and correct answers remained on the screen while the next problem appeared (up to four problems at a time). This allowed some spontaneous comparison across problems during the provision of feedback, much like a summative answer key. In the immediate-feedback condition, the problems and answers disappeared before the next problem.

In both feedback conditions, the feedback was based solely on the child's numerical answer and did not depend on the strategy reported. On a few occasions, children reported using a correct strategy, but obtained an incorrect answer due to an arithmetic mistake and received negative feedback. However, these mismatches were rare (7% of all trials) and exclusion of children who experienced more than two mismatches ($n = 7$) did not impact the results.

Data Analysis

Two children were missing data on the equation-encoding section of the screening measure and two children failed to provide their date of birth and were missing values for their age. Imputing missing independent variables leads to more precise and unbiased conclusions than omitting participants with missing data (Peugh & Enders, 2004). I used the expectation-maximization algorithm for maximum likelihood estimation via the missing values analysis in SPSS (Schafer & Graham, 2002) to impute the missing encoding scores and ages.

To examine children's performance on the primary outcome measures, I performed a series of ANCOVAs with condition as a between-subject variable. Specifically, condition was dummy coded with immediate feedback and summative feedback entered into the models, and no feedback as the reference group. On several measures, scores were not normally distributed. In those cases, I used binomial logistic regression. Again, condition was dummy coded. In all models, I included children's age and their score on the screening measure as covariates. Preliminary analyses revealed no interactions with age or screening measure scores so these interaction terms were not retained in the final models. I also explored the impact of grade on learning outcomes by including it in the models. There were rarely any differences by grade and grade never interacted with feedback. Thus, grade was not retained in the final models.

Results

Screening Measure

On average, children in the final sample solved 1.0 (*SD* = 1.0) problem correctly (out of 5), encoded 1.7 (*SD* = 1.2) problems correctly (out of 4), and only 6% of children provided a

relational definition of the equal sign. Performance on the three tasks did not differ as a function of condition, ps > .55. I created a composite measure of children's performance by summing z-scores across the three tasks. Composite scores ranged from –2.62 to 6.95 (M = 0.00, SD = 1.97) and did not differ as a function of condition, p = .91.

Knowledge Check

The knowledge induction was largely successful. Most children (85%) exhibited knowledge of a correct strategy on the first problem following instruction, 7% needed a second problem, and 2% were successful by the fifth attempt. The remaining 6% of children never used a correct strategy after five attempts with repeated instruction after each problem. Difficulties were often due to poor arithmetic fact knowledge and weak counting skills. For these children, the experiment was stopped and remedial tutoring was provided, as they were clearly not ready to learn about or solve these problems. This resulted in a sample of 95 children (no-feedback, n = 32; immediate-feedback, n = 32; summative-feedback, n = 31).

Intervention Measures

As in Experiment 1, for analyses of children's verbal strategy reports, I focused on the 9 math equivalence problems. The pattern of results was similar if I considered all 12 problems.

Correct Strategy Use. The frequency of correct strategy use during the intervention was similar for children in the no-feedback (M = 88%, SE = 4%), immediate-feedback (M = 85%, SE = 4%), and summative-feedback conditions (M = 82%, SE = 4%). As in Experiment 1, there was no significant difference between the immediate-feedback and no-feedback conditions, p = .53. There was also no effect of summative feedback relative to no feedback, p = .24. A follow-up analysis revealed no significant difference between the two feedback types, p = .58.

Correct Strategy Generation. The number of different types of correct strategies used was also similar for children in the no-feedback (M = 1.4, SE = 0.1), immediate-feedback (M = 1.3, SE = 0.1), and summative-feedback conditions (M = 1.2, SE = 0.1). As in Experiment 1, there was no significant effect of immediate feedback relative to no feedback, p = .78. There was also no effect of summative feedback relative to no feedback, p = .12. A follow-up analysis revealed no significant difference between the two feedback types, p = .20.

Incorrect Strategy Generation. The number of different types of incorrect strategies used was highest with immediate feedback (M = 1.0, SE = 0.2), next highest with summative feedback (M = 0.8, SE = 0.2), and lowest with no feedback (M = 0.4, SE = 0.2). As in Experiment 1, there was a significant effect of immediate feedback relative to no feedback, $F(1, 90)$ = 5.31, p = .02, η_p^2 = .06. Unexpectedly, there was also a marginal effect of summative feedback relative to no feedback even though the summative feedback had not yet occurred during this phase of the experiment, $F(1, 90)$ = 3.64, p = .06, η_p^2 = .04. A follow-up analysis revealed no significant difference between the two feedback types, p = .69. Only two children perseverated and used the same incorrect strategy across problems, and they were both in the no-feedback condition.

Positive vs. Negative Feedback. The majority of children in the feedback conditions received a mix of positive (i.e., that's correct) and negative (i.e., that's incorrect) feedback. A small portion of children (n = 8 in immediate-feedback and n = 2 in summative-feedback) received positive feedback on all trials, but no child received negative feedback on all trials.

Cognitive Load. Children's ratings of cognitive load were similar across conditions, but somewhat higher with summative feedback (M = 2.3, SE = 0.1) than with immediate feedback (M = 2.1, SE = 0.1) or no feedback (M = 2.0, SE = 0.1). As in Experiment 1, there was no effect of immediate feedback relative to no feedback, p = .38. There was a marginal effect of

summative feedback relative to no feedback, $F(1, 89) = 2.98$, $p = .09$, $\eta_p^2 = .03$. A follow-up analysis revealed no significant difference between the two feedback types, $p = .40$.

Self-Assessment. The percent of children scoring a 4 out of 4 (i.e., reporting positive self-assessment on each of the four items) was similar across the no-feedback (87%), immediate-feedback (78%), and summative-feedback conditions (84%). A binomial logistic regression showed that there were no significant effects of immediate feedback relative to no feedback, $p = .23$, or summative feedback relative to no feedback, $p = .95$. A follow-up analysis revealed no significant difference between the two feedback types, $p = .24$.

Intervention Summary. As in Experiment 1, feedback primarily had neutral effects for children with knowledge of a correct strategy during the intervention. Specifically, immediate correct-answer feedback did not impact the frequency of correct strategy use, it facilitated the generation of more incorrect strategies, and it led to similar ratings of task difficulty relative to no feedback. As expected, the summative-feedback condition was not reliably different from the no-feedback condition on any measure during the intervention.

Posttest Procedural Knowledge

Children's percent correct on procedural knowledge at posttest was highest with no feedback ($M = 83\%$, $SE = 5\%$), lower with immediate feedback ($M = 78\%$, $SE = 5\%$), and lowest with summative feedback ($M = 71\%$, $SD = 5\%$). There was no significant effect of immediate feedback relative to no feedback, $p = .52$. There was a marginal effect of summative feedback relative to no feedback, $F(1, 90) = 3.60$, $p = .06$, $\eta_p^2 = .04$. A follow-up analysis revealed no significant difference between the two feedback types, $p = .22$.

However, scores on procedural knowledge at posttest were high and not normally distributed. Across conditions, the vast majority of children (82%) solved 50% correct or higher

with a full 40% of children solving all of the items correctly. I used binomial logistic regression to predict the log of the odds of scoring 100% correct. The results are displayed in Figure 3.
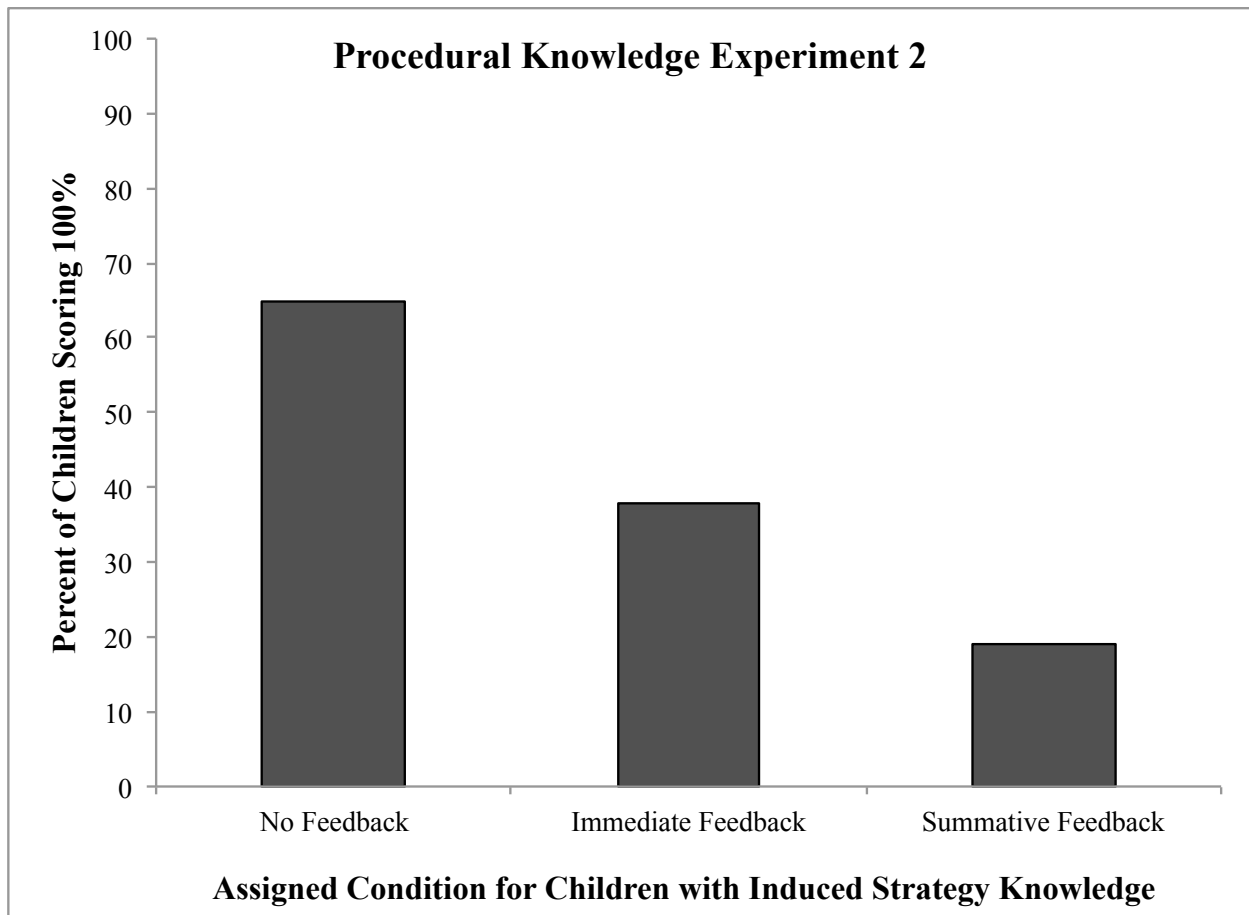
Figure 3. Procedural knowledge at posttest in Experiment 2

As in Experiment 1, there was a significant, negative effect of immediate feedback. Children in the immediate-feedback condition were *less* likely than children in the no-feedback condition to score 100% on the posttest (38% vs. 65%), $\beta = -1.16$, $z = 2.17$, *Wald* $(1, N = 95) = 4.70$, $p = .03$, $OR = 0.31$. There was also a significant, negative effect of summative feedback. Children in the summative-feedback condition were also less likely than children in the no-feedback condition to score 100% on the posttest (19% vs. 65%), $\beta = -2.07$, $z = 3.48$, *Wald* $(1, N = 95) = 12.10$, $p = .001$, $OR = 0.13$. A follow-up analysis revealed no significant differences

between the two feedback conditions, *p* = .13. Exploratory analyses revealed that the results remain unchanged after excluding children who received only positive feedback. Further, the pattern of results was consistent across the learning and transfer items, though more pronounced for transfer. Overall, children with knowledge of a correct strategy learned more from no feedback than immediate- or summative-feedback. This difference was consistent whether we examined percent correct or percent of children at mastery, though more reliable with the latter.

As in Experiment 1, the conceptual knowledge results are reported in full in Appendix C. Consistent with the procedural knowledge results, feedback (both immediate and summative) had negative effects on conceptual knowledge for children with knowledge of a correct strategy.

Discussion

Experiment 2 was largely consistent with Experiment 1 and supported my first hypothesis. Children with induced knowledge of a correct strategy benefitted more from no feedback than from immediate feedback. The negative effect of immediate correct-answer feedback occurred on children's procedural and conceptual knowledge at posttest. Further, the intervention results were similar to those from Experiment 1 and provided further insight into the effects of feedback during learning. Specifically, immediate feedback facilitated the generation of incorrect strategies relative to the no feedback condition, but had no impact on correct strategy generation. Immediate feedback did not have a reliable impact on reported cognitive load or self-assessment, failing to support the potential role of these factors in explaining the negative effects of feedback. In Experiment 1, immediate feedback marginally reduced correct strategy use during the intervention. Here, the direction of the effect was similar, but not as strong.

39

Contrary to my second hypothesis, the effects of summative feedback were also negative. Specifically, children who received summative feedback exhibited lower procedural and conceptual knowledge at posttest than children who received no feedback. Indeed, the two feedback types (immediate and summative) did not differ significantly from one another on the posttest. One possibility is that both types of feedback trigger cognitive or affective responses that interfere with learning. Thus, *whether* these responses are elicited may be more important than *when* they are elicited (e.g., during or after problem solving). This may be particularly true when knowledge is assessed immediately as it was on the posttest in the current experiment.

In general, the results from Experiments 1 and 2 confirm that feedback can have negative effects that are robust across various feedback types (i.e., verification vs. correct-answer and immediate vs. summative). However, several limitations restrict the conclusions that can be made. For example, Experiments 1 and 2 were both conducted in one-on-one tutoring settings with a novel experimenter/tutor. Feedback may have especially powerful effects in settings with enhanced individualized attention and with a salient other-person presence.

To better understand the potential negative effects of immediate and summative feedback, a third experiment was conducted similar to Experiment 2, but in a classroom setting. I wanted to verify whether the negative effects of feedback generalized to a more typical learning environment. In a classroom, it is typical to withhold feedback until all students are finished with a set of problems. I also wanted to examine the effects of feedback in a setting with reduced individualized attention from an authority figure. In a one-on-one setting, the tutor is clearly aware of a student's answers and whether they are correct or incorrect, potentially focusing attention on the self as opposed to the task. In a classroom setting, individualized attention from the instructor is reduced, which may attenuate any negative effects of feedback that result from

drawing attention to one's self. Although, the presence of peers may also draw attention to the

self as opposed to the task and enhance the evaluative component of the task. Thus, I conducted

Experiment 3 to explore how feedback impacts learning in the classroom.

CHAPTER IV

EXPERIMENT 3

The goal of Experiment 3 was to provide additional insight into the presence and timing

of feedback for learners with some prior knowledge. Again, I focused on children with some

prior knowledge for whom feedback had negative effects relative to no feedback. Thus, I used a

pre-familiarization technique to ensure that all children had knowledge of a correct strategy.

This study was similar to Experiment 2, but it was conducted in a classroom setting rather

than in a one-on-one tutoring setting. Specifically, we worked with children in small groups

within their classroom in one of three conditions. Children assigned to the same condition

formed one group. Because some students in a group solved a problem faster than other students

in the group, immediate feedback did not always occur right after a student solved the problem.

Thus, the primary distinction between immediate feedback and summative feedback in this

experiment is whether it was provided after *each* problem or after *all* of the problems.

In line with Experiments 1 and 2, I predicted that immediate feedback would have

negative effects relative to no feedback. Because there is less individualized attention from the

instructor and thus, potentially less interference from affective responses, the negative effects of

immediate feedback may be attenuated in the classroom setting. Although, more attention from

peers may offset the reduced attention from the instructor and thus, maintain the negative effects

of feedback found in previous experiments. My predictions for the effects of summative

feedback were tentative. On the one hand, I expected summative feedback to have positive

effects relative to no feedback because it reduces interference *during* problem solving and has

been shown to facilitate retention. On the other hand, results from Experiment 2 indicate that summative feedback may still have negative effects relative to no feedback. The feedback may still produce cognitive or affective responses that change the strength of one's knowledge and thus influence application of the just-learned knowledge on future assessments.

Method

Participants

All second- and third-grade classes from one public and one private elementary school were invited to participate, with a total of 274 children completing the initial screening measure. These children were from 14 different classrooms, ten of which were in the public school. The public school served predominantly middle-class White students in kindergarten through fifth grade (81% White, 10% Black, 5% Hispanic, 2% Asian), with a minority (29%) receiving free or reduced priced lunch. In 2014, 66% of students in third- and fourth-grade scored proficient or higher on the TCAP math assessment. The private school served primarily middle-class White students in pre-K through eight grade (78% White, 6% Black, 6% Hispanic, 6% Asian).

All children were included in the study, regardless of their performance on the screening measure, as is typical in classroom settings. However, data from 31 children were excluded from the final analyses for failing to complete all intervention activities. Most of these children ($n = 20$) were from a single classroom in which we were unable to finish the lesson and posttest in the allotted time. Seven children were absent the day of the intervention, two children left early from the intervention, and two skipped key activities. The final sample contained 243 children (*M* age = 8.3 yrs, *min* = 6.6 yrs, *max* = 10.2 yrs; 107 girls, 136 boys).

Design

The study had a between-subjects design. The intervention occurred in children's classrooms during their normal mathematics instruction. Within each classroom, children were randomly assigned to conditions: (1) no feedback ($n = 80$), (2) immediate feedback ($n = 82$), and (3) summative feedback ($n = 81$). Research assistants worked with small groups of children assigned to the same condition, allowing children within the same classroom to be assigned to different conditions. There were no differences between conditions in terms of age, grade, gender, or ethnic minority status, ps > .2.

Materials

The materials were identical to those in Experiment 1 with a few exceptions. First, I administered a brief version of the conceptual knowledge posttest that contained five items rather than ten (see Appendix C). Second, to assess children's cognitive load, I only administered the validated three-item task difficulty scale, as in Experiment 2. Third, for the strategy instruction, the example problems were printed on large, laminated paper and shown to the group, rather than presented on a computer screen. Fourth, for problem solving, each child received a problem-solving packet with problems printed one per page, rather than viewing them on a computer.

Procedure

The procedure was identical to Experiment 2, but adapted to be suitable in a classroom setting. For the intervention, children assigned to the same condition sat in a small group in their classroom. Each of the three groups in the classroom was supervised by a different research assistant. The average group size was 6 children ($SD = 1.0$), but ranged from 5 to 9 depending on the number of children in the class. For the knowledge check, all children solved two math equivalence problems on their own, rather than tailoring the number of problems to children's

individual performance. After all children in the group had completed the first knowledge-check problem, general instruction on the taught strategy was repeated without revealing the correct answer, and they were all asked to solve a second problem.

For the problem-solving phase, in the *no-feedback* condition, children worked on the problems at their own pace and did not receive any feedback on their answers. Once finished, they were asked to go over the problems on their own and check their work using a purple pen to better equate time-on-task across conditions. In the *immediate-feedback* condition, children worked on the problems one at a time and received trial-by-trial correct-answer feedback after each problem. When all children in the small group had solved each problem, the research assistant read the problem out loud along with the correct answer. Children then marked in their packets whether their answer was correct (a check mark) or incorrect (an x). In the *summative-feedback* condition, children worked on the problems at their own pace and received correct-answer feedback after all 12 problems had been solved. When all children in the small group had solved all problems, the research assistant read the problems out loud along with the correct answers. Children marked in their packets whether their answers were correct or incorrect.

Due to the group setting, children did not provide verbal strategy reports on each problem. However, we coded their problem-solving strategies based on their numerical answers and written work (e.g., for $3 + 7 = \_\_ + 6$, an answer of 16 indicated an incorrect "add-all" strategy and an answer of 4 indicated a correct strategy.) As in prior work (McNeil, 2008), responses within +/-1 of the correct answer were coded as reflecting a correct strategy. A second rater coded 30% of the responses and interrater agreement was high (kappa = .95).

Following the problem solving, children completed the cognitive load measure. The items were printed on the back of the problem-solving packet and children circled their

responses. Next, children completed the immediate posttest on their own at their own pace. Approximately one week later (*M* = 6.6 days, *SD* = 0.9, *min* = 5, *max* = 9), children completed a retention test in their classrooms.

Data Analysis

We initially worked with children in 42 small groups of 5–9 children each. To test for nonindependence, I calculated unconditional intraclass correlations on the outcomes, using the approach that allows for negative nonindependence (Kenny, Kashy, Mannetti, Pierro, & Livi, 2002). The intraclass correlations were moderate with values of .16 for intervention problem solving and .17 for posttest procedural knowledge. Because there was nonindependence in the data, I used multilevel modeling to account for this nesting within group. Specifically, all models had two levels: (1) the individual level and (2) the small-group level.

For most outcome measures, I used linear regression models. I specified the use of restricted maximum-likelihood estimation and compound symmetry for the variance-covariance structure in the models (Kenny, Kashy, & Cook, 2006). For outcomes that were not normally distributed, I used a generalized linear mixed model, specifying the response distribution as Bernoulli with a logit link (Snijders & Bosker, 2012). These models allow for the analysis of dichotomous outcomes and predict the log of the odds of a student scoring 1 versus 0.

As in Experiment 2, condition was dummy coded with immediate feedback and summative feedback entered into the models, and no feedback as the reference group. Screening measure scores and age were included as covariates. Both variables were mean-centered. Exploratory analyses revealed no interactions with screening measures scores or age so these terms were not included in the final models. I also explored the impact of grade on learning outcomes by including it in the models. There were general trends indicating that third-grade

children often scored somewhat higher than second-grade children, but these main effects of grade rarely reached statistical significance and grade did not reliably interact with feedback across outcome measures. Thus, grade was not retained in the final models.

Results

Screening Measure

On average, children in the final sample solved 1.0 ($SD = 1.6$) math equivalence problem correctly (out of 4), encoded 2.1 ($SD = 1.4$) problems correctly (out of 4), and 16% of children provided a relational definition of the equal sign. Performance on the three tasks did not differ as a function of condition, ps $> .15$. We created a composite measure of children's performance by summing z-scores across the three tasks. Composite scores ranged from -2.57 to 5.50 ($M = 0.00$, $SD = 1.96$). The scores did not differ as a function of condition, $p = .61$.

Knowledge Check

The knowledge induction was largely successful. The vast majority of children exhibited knowledge of a correct strategy after the instruction by solving both ($n = 224$ children) or one ($n = 8$ children) of the knowledge check problems correctly. The remaining 11 children solved both problems incorrectly and were excluded from subsequent analyses as their performance indicated they likely did not know a correct strategy. This resulted in a final sample of 232 children (no-feedback, $n = 77$; immediate-feedback, $n = 80$; summative-feedback, $n = 75$). The pattern of findings was the same if these eleven children were included.

Intervention Measures

As in Experiments 1 and 2, for analyses of strategy use, I focused on the 9 math equivalence problems. But, the pattern of results was similar if I considered all 12 problems.

Correct Strategy Use. In contrast to Experiments 1 and 2, the frequency of correct strategy use during the intervention was highest for children in the immediate-feedback condition ($M = 92\%$, $SE = 2\%$), lower in the summative-feedback condition ($M = 88\%$, $SE = 2\%$), and lowest in the no-feedback condition ($M = 83\%$, $SE = 3\%$). There was a significant effect of immediate feedback relative to no feedback, $\beta = 7.50$, $p = .04$. There was no effect of summative feedback relative to no feedback, $\beta = 5.28$, $p = .14$. A follow-up analysis revealed no significant difference between the two feedback types, $\beta = 2.22$, $p = .53$.

Strategy Generation. Strategy generation was difficult to examine in this classroom study. Based on children's written work, it was difficult to discern the types of correct strategies children used. Further, although children's errors were informative for discerning the types of incorrect strategies children used, they were rare. Indeed, the average number of different types of incorrect strategies used was less than one in the immediate-feedback ($M = 0.6$, $SE = 0.1$), summative-feedback ($M = 0.6$, $SE = 0.1$), and no-feedback conditions ($M = 0.8$, $SE = 0.1$). Thus, unlike Experiments 1 and 2, incorrect strategy generation did not seem to play a role.

Cognitive Load. As in Experiments 1 and 2, children's ratings of cognitive load were similar in the immediate- ($M = 1.7$ out of 4, $SE = 0.1$), summative- ($M = 1.7$, $SE = 0.1$) and no-feedback ($M = 1.8$, $SE = 0.1$) conditions. There were no significant condition effects, $ps > .87$.

Posttest Procedural Knowledge

Children's percent correct on the procedural knowledge scale at posttest was similar in the immediate- ($M = 83\%$, $SD = 26\%$), summative- ($M = 83\%$, $SD = 23\%$), and no-feedback (M

= 80%, *SD* = 29%) conditions. In a multi-level linear regression model, there were no condition effects, *p*s > .55. However, consistent with Experiment 2, scores on the procedural knowledge scale at posttest were high and not normally distributed. Across conditions, the vast majority of children (89%) solved 50% correct or higher, with a full 47% of children solving all the items correctly. I ran a generalized linear mixed model to predict the log of the odds of scoring 100% correct on the posttest. The percent of children scoring 100% was similar in the immediate- (47%), summative- (48%), and no-feedback (46%) conditions. Again, there were no condition effects, *p*s > .90. Overall at posttest, feedback had a neutral effect relative to no-feedback.

Retention Test Procedural Knowledge

At retention test, children's percent correct on procedural knowledge was similar in the immediate- (*M* = 77%, *SD* = 30%), summative- (*M* = 75%, *SD* = 32%), and no-feedback (*M* = 75%, *SD* = 35%) conditions. In a multi-level linear regression model, there were no condition effects, *p*s > .65. As on the posttest, however, scores at retention were high and not normally distributed. Across conditions, the vast majority of children (75%) solved 50% correct or higher, with a full 45% of children solving all of the items correctly. I ran a generalized linear mixed model to predict the log of the odds of scoring 100% correct on the retention test. There was a significant, negative effect of immediate feedback. Children in the immediate-feedback condition were *less* likely than children in the no-feedback condition to score 100% on the retention test (41% vs. 53%), β = –0.77, *p* = .01, *OR* = 0.46. There was also a significant, negative effect of summative feedback. Children in the summative-feedback condition were also less likely than children in the no-feedback condition to score 100% on the posttest (42% vs. 53%), β = –0.71, *p* = .02, *OR* = 0.49. A follow-up analysis revealed no significant differences between the two feedback conditions, *p* = .82. Feedback had a negative effect for children reaching mastery.

Overall at retention, the conclusions depended on the outcome measure. Focusing on

percent correct, feedback had a neutral effect. Focusing on mastery (percent of children solving

100%), feedback had a negative effect, such that fewer children in the feedback conditions reach

mastery compared to the no-feedback condition. A qualitative look at the data suggests that this

discrepancy is due to the different distributions of scores across conditions (see Figure 4).

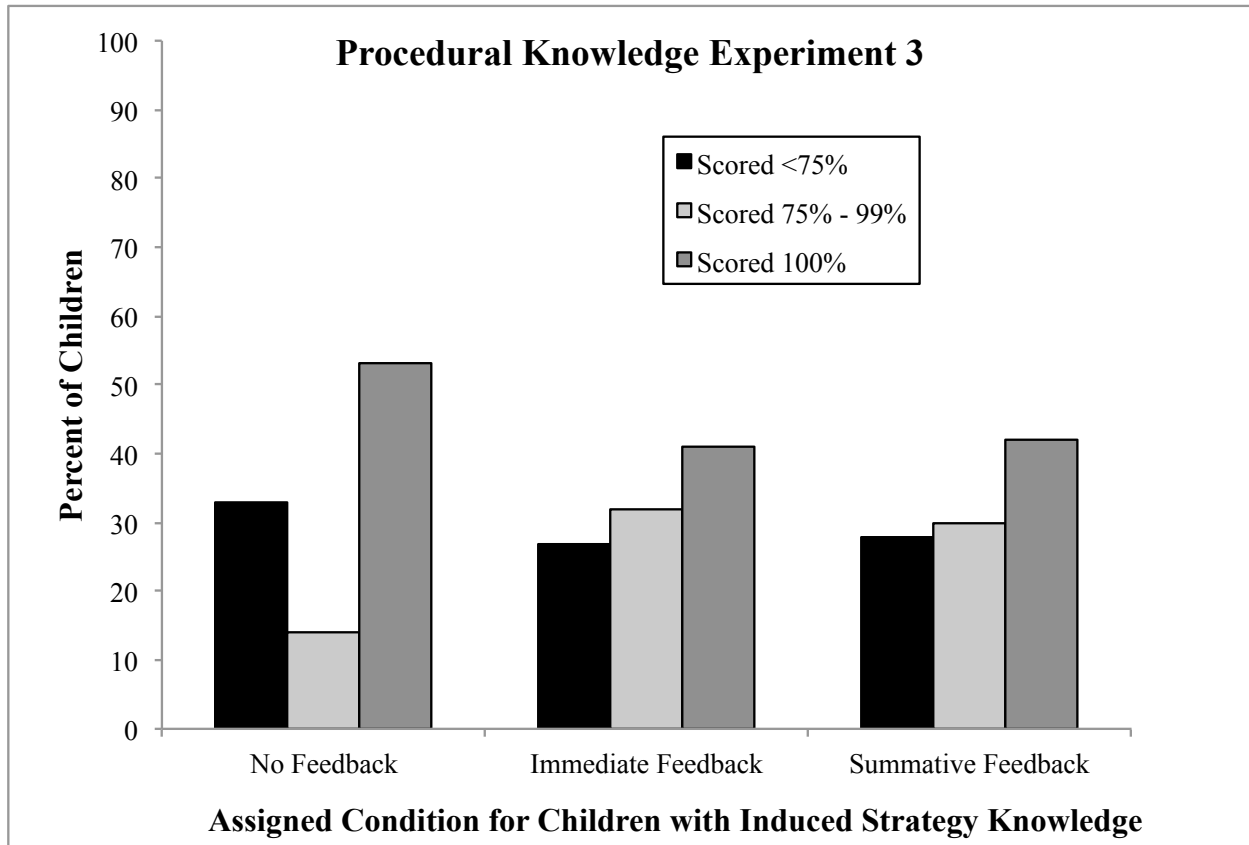*Figure 4. Procedural knowledge at retention test in Experiment 3*

Children in the no-feedback condition tended to score really high or relatively low

compared to the rest of the sample. However, relatively more children in the two feedback

conditions tended to score moderately high. Thus, fewer children in the feedback conditions

reached mastery, but fewer remained at relatively low levels of knowledge compared to the no-

feedback condition. Further, exploratory analyses suggest the negative effects of feedback were primarily present in third-graders (not second-graders), which is consistent with the distribution data as more third-grade children were at mastery (scored 100%) than second-grade children.

As in Experiments 1 and 2, the conceptual knowledge results are reported in full in Appendix C. Neither type of feedback (immediate and summative) reliably impacted children's conceptual knowledge at the posttest or retention test.

Discussion

The general conclusion from Experiment 3 is consistent with Experiments 1 and 2. Feedback, whether presented immediately or after a delay, can have negative effects relative to no feedback on mathematics learning for children with higher prior knowledge. However, the specific results were somewhat at odds with those from the previous experiments. In Experiments 1 and 2, immediate feedback had neutral (and trending negative) effects on performance during the intervention. Here, immediate feedback had positive effects during the intervention, resulting in higher use of correct strategies relative to no feedback and summative feedback. Further, the negative effects of feedback in Experiments 1 and 2 emerged on the posttest assessment (and diminished on the retention test in Experiment 1). Here, the negative effects of feedback were not present on the posttest, but emerged one week later, and only for children at mastery (scoring 100% on the retention test). Thus, Experiment 3 suggests that the negative effects of correct-answer feedback during problem solving generalize to a classroom setting with a novel instructor (rather than the regular, routine teacher). However, the negative effects may emerge later than in a one-on-one tutoring setting and may be less robust.

The effects of immediate feedback in this experiment mirror some found in prior work. Specifically, some researchers have found that immediate feedback can benefit performance during initial learning, but result in increased errors on delayed retention tests (e.g., Schmidt, Young, Swinnen, & Shapiro, 1989, see also Schmidt & Bjork, 1992). Here, immediate feedback boosted problem-solving performance on the intervention trials, but resulted in lower mastery on the one-week retention test. One idea put forth is that learners come to rely on the immediate presentation of feedback and flounder on subsequent assessments when it is no longer available. In this scenario, a no-feedback condition may represent a "desirable difficulty" (Bjork, 1994). That is, it creates a difficult initial learning environment as there is little guidance on how to detect and correct errors. However, it is desirable in the sense that it triggers deep processing of the problems in a way that enhances later retention (Bjork & Bjork, 1992).

However, several previous experiments suggest that delayed feedback may be particularly beneficial for retention of knowledge (e.g., Butler et al., 2007; Schmidt et al., 1989). That was not the case here. In fact, across Experiments 2 and 3, the timing of feedback did not matter. Immediate feedback and summative feedback produced similar results, suggesting the mere provision of feedback may matter more than when it is provided during mathematics problem solving. Although far from definitive, these results suggest that interference *during* initial problem solving is not a major reason for the negative effects of immediate feedback (given that summative feedback does not occur until after all the problems have been solved).

One thing that immediate and summative feedback have in common is the potential to focus attention on children's self-image and abilities at the expense of learning. For example, higher-knowledge learners who receive negative feedback (i.e., that's incorrect), may lose confidence in their abilities and focus their attention on how well (or not well) they are doing

52

rather than focusing on learning from the task. Attention on the self may help explain why the negative effects of feedback were less robust in this classroom experiment. In a classroom, there is less individualized attention on each student from the teacher, and thus, potentially less interference from affective responses that arise from the presence of an authority figure. However, there is more attention from *peers* relative to a one-on-setting, which could actually enhance children's awareness of their performance and evaluations, particularly for children who take more time or score lower on tasks relative to their classmates. Thus, the role of attention on the self is ambiguous in a classroom setting and warrants further, systematic investigation. For example, feedback may be less likely to have negative effects if provided by an existing teacher who has already established trusting relationships with her students.

To better understand the role of attention from an authority figure, I conducted a fourth experiment similar to Experiments 2 and 3 in a one-on-one tutoring context. A key change was the removal of the experimenter during the problem-solving task. Rather than receiving feedback from a person, children received feedback from a computer. Computer-feedback is more likely to focus attention on the task as opposed to one's self-image (Kluger & DeNisi, 1996), and it is often viewed as a less evaluative source of information. For example, in one study, subjects had the option of requesting corrective information after negative feedback trials (Karabenick & Knapp, 1988). While 86% of subjects asked for information from the computer, only 36% asked for information when they believed a person was providing help via a computer network. Some evidence suggests that computer-feedback is more effective than person-feedback for adults (Kluger & Adler, 1993), but research with younger children has found no differences between feedback from a person and feedback from a computer (see Barringer & Gholson, 1979).

A second key change was the delay of the posttest until the following day. The goal was to allow time for any affective responses (e.g., ego-threat) that resulted from feedback to dissipate before testing. Unlike the immediate posttests in the previous experiments, the delayed posttest may less associated with the problem-solving task during the intervention. Further, given the variety of tasks that will likely occur between the receipt of feedback and the next-day assessment, children may be better able to focus their attention on the task at hand.

CHAPTER V

EXPERIMENT 4

The goal of Experiment 4 was to provide additional insight into the presence and timing of feedback for learners with some prior knowledge. As in Experiments 2 and 3, I used a pre-familiarization technique to ensure that all children had knowledge of a correct strategy. Then, during the problem-solving task, children received no feedback, immediate feedback, or summative feedback. This experiment occurred in a one-on-one tutoring context.

This study was similar to Experiments 2 and 3, but with key changes to reduce the potential effects of drawing attention to the child's performance and the potential role of ego-threat. First, during the problem-solving task, the experimenter presence was removed so that the feedback was provided solely by the computer. Second, the feedback did not contain explicit right/wrong verification. It only contained the correct answer. Third, the posttest was provided the following day rather than immediately following the intervention to allow for any affective responses that might be induced by feedback to dissipate by the time of testing.

If drawing attention to self and ego-threat are *not* the primary reason for feedback's negative effects for higher knowledge learners, then I expect to replicate the negative effects of feedback found in previous experiments. However, if drawing attention to self and ego-threat are the primary reason for feedback's negative effects, then I expect both types of feedback in this experiment to have positive effects relative to no-feedback on the posttest assessment.

Method

Participants

Initial participants were 88 second-grade children, none of whom participated in Experiments 1 through 3. Children were recruited from the one public and one private school. The public school served ethnically-diverse students in kindergarten through fourth grade (60% White, 31% Black, 5% Hispanic, 4% Asian), just under half of whom (41%) received free or reduced priced lunch. In 2014, 52% of students in third- and fourth-grade scored proficient or higher on the TCAP math assessment. The large, private school primarily served middle- to upper-class White students in pre-kindergarten through eighth grade with three classes per grade.

Of the initial 88 children, 77 (88%) met criteria for participation because they scored below 80% on a problem-solving screening measure. Data from 2 additional children were excluded (one asked to stop halfway through the intervention, and one was due to experimenter error in administering the intervention). The final sample contained 75 second-grade children (*M* age = 8.2 yrs, *min* = 7.4 yrs, *max* = 9.2 yrs; 41 girls, 34 boys).

Design

The study had a between-subjects design with children randomly assigned to conditions: (1) no feedback (*n* = 24), (2) immediate feedback (*n* = 25), and (3) summative feedback (*n* = 26). There were no differences between conditions in terms of age or gender, *p*s > .40.

Materials

The materials were identical to those in Experiment 2 with two exceptions. First, during the initial instruction, all four example problems had the blank at the end of the math equivalence problem (e.g., 3 + 4 + 2 = 3 + __). In the previous experiments, I used two problem structures to

increase the generalizability of the strategy (blank at the end and blank right after the equal sign).

Here, I narrowed the instruction to focus on a single problem structure to enhance the challenge

of the subsequent problem-solving task and the variability in performance. Second, I changed the

response format for the self-assessment items to increase the variation in scores. Rather than

selecting from a dichotomous option (good student vs. not good student), children rated their

agreement on a four point scale ranging from *strongly disagree* to *strongly agree*. The statements

were in reference to the positive attribute: "The problem-solving task made me feel like I

was…[good at solving the problems, a good student, a nice student, or a smart student]."

Internal consistency was good ($\alpha = .80$), but the range of responses was still somewhat restricted

as very few children ever selected a 1 (strongly disagree) or 2 (disagree).

Procedure

The procedure was identical to Experiment 2 with a few exceptions. First, during the

problem-solving task, the experimenter presence was removed. Specifically, children were told

they would work on the computer by themselves for this portion of the session so that they could

work at their own pace and not worry about the experimenter. Then, the experimenter sat a short

distance away and engaged in a different task (e.g., read a book) until the child had completed

the problem-solving task. Due to the removal of the experimenter, children did not provide

verbal strategy reports on each problem. However, we coded their problem-solving strategies

based on their numerical answers. A second rater coded 30% of the responses and interrater

agreement was high (kappa = .97). Second, the feedback provided by the computer did not

contain right/wrong verification, it only contained the correct answer (e.g., "10 is the correct

answer"). Although the right/wrong judgment was implicit (via comparison of the child's answer

with the correct answer), there was no explicit signal provided from the computer (e.g., check

mark, noise, etc.). Third, the posttest was administered the next day, rather than immediately after the intervention. Children were pulled out of the classroom in small groups and completed the posttest independently. Due to absences on the day of the scheduled posttest, three children completed the posttest more than one day later (i.e., 2, 4, and 5 days later).

Data Analysis

To examine children's performance on the primary outcome measures, I performed a series of ANCOVAs with condition as a between-subject variable. Specifically, condition was dummy coded with immediate feedback and summative feedback entered into the models, and no feedback as the reference group. On one measure, scores were not normally distributed. In that case, I used binomial logistic regression. In all models, I included children's age and their score on the screening measure as covariates. Preliminary analyses revealed no interactions with age and only two interactions with screening measures. Specifically, for posttest procedural knowledge, both feedback conditions interacted with screening measure scores. The direction of effects was similar for all children, but stronger for children with low screening measure scores. Because interactions with age and screening measure scores were rare and did not represent changes in the direction of effects, these interaction terms were not retained in the final models.

Results

Screening Measure

On average, children in the final sample solved 1.3 ($SD = 1.2$) problems correctly (out of 5), encoded 1.8 ($SD = 1.4$) problems correctly (out of 4), and only 13% of children provided a relational definition of the equal sign. Performance on the three tasks did not differ as a function

of condition, $ps > .48$. I created a composite measure of children's performance by summing z-scores across the three tasks. Composite scores ranged from –2.70 to 6.23 ($M = 0.00$, $SD = 2.10$) and did not differ as a function of condition, $p = .70$.

Knowledge Check

The knowledge induction was largely successful. Most children (88%) exhibited knowledge of a correct strategy on the first problem following instruction, 7% needed a second problem, and 3% were successful by the third attempt. The remaining 3% of children ($n = 2$) never used a correct strategy after three attempts with repeated instruction after each problem. For these children, the experiment was stopped and remedial tutoring was provided, as they were clearly not ready to learn about or solve these problems. This resulted in a sample of 73 children (no-feedback, $n = 24$; immediate-feedback, $n = 24$; summative-feedback, $n = 25$).

Intervention Measures

For analyses of children's strategy use, I focused on the 9 math equivalence problems. The pattern of results was similar if I considered all 12 problems.

Correct Strategy Use. The frequency of correct strategy use during the intervention was similar for children in the immediate- ($M = 77\%$, $SE = 5\%$) and summative-feedback ($M = 79\%$, $SE = 6\%$) conditions, though somewhat lower in the no-feedback condition ($M = 70\%$, $SE = 6\%$). As in Experiments 1 and 2, there were no reliable condition effects. There was no significant difference between the immediate-feedback and no-feedback conditions, $p = .23$. There was a marginal effect of summative feedback relative to no feedback $F(1, 68) = 3.24$, $p = .08$, $\eta_p^2 = .03$, even though the summative feedback had not yet occurred during this phase of the experiment. A follow-up analysis revealed no difference between the two feedback types, $p = .57$.

Strategy Generation. It was difficult to discern the types of correct strategies children used. However, children's errors were informative for discerning the types of incorrect strategies children used. The average number of different types of incorrect strategies used was similar in the immediate-feedback ($M = 1.3$, $SE = 0.2$), summative-feedback ($M = 1.2$, $SE = 0.3$), and no-feedback conditions ($M = 1.3$, $SE = 0.3$). Only two children perseverated and used the same incorrect strategy across problems, and they were both in the no-feedback condition. Thus, unlike Experiments 1 and 2, incorrect strategy generation did not seem to play a large role.

Positive vs. Negative Feedback. The majority of children in the feedback conditions received a mix of positive (i.e., correct) and negative (i.e., incorrect) feedback. One child in the summative-feedback condition received only negative feedback, and a few children ($n = 6$ in immediate-feedback and $n = 8$ in summative-feedback) received only positive feedback.

Cognitive Load. As in previous experiments, children's ratings of cognitive load were similar for children in the immediate- ($M = 1.9$, $SE = 0.1$), summative- ($M = 1.9$, $SE = 0.1$), and no-feedback ($M = 1.8$, $SE = 0.1$) conditions. There were no effect of immediate feedback relative to no feedback, $p = .82$, and no effect of summative feedback relative to no feedback, $p = .89$. A follow-up analysis revealed no significant difference between the two feedback types, $p = .92$.

Self-Assessment. Children's ratings of positive self-assessment were similar in the immediate- ($M = 3.5$, $SE = 0.1$) and no-feedback ($M = 3.4$, $SE = 0.1$) conditions, but lower in the summative-feedback condition ($M = 3.1$, $SE = 0.1$). There was no effect of immediate-feedback relative to no-feedback, $p = .80$. There was a significant effect of summative feedback relative to no feedback $F(1, 67) = 4.99$, $p = .03$, $\eta_p^2 = .07$. A follow-up analysis revealed a significant difference between the two feedback types, $F(1, 67) = 6.39$, $p = .01$, $\eta_p^2 = .09$.

Intervention Summary. As in Experiments 1 and 2, feedback primarily had neutral effects for children with knowledge of a correct strategy during the intervention. Specifically, immediate feedback did not impact correct strategy use or the generation of incorrect strategies, and it led to similar ratings of task difficulty and self-assessment relative to no feedback. Unexpectedly, children in the summative-feedback condition had slightly higher correct strategy use and lower ratings of positive self-assessment relative to the other two conditions. Receiving feedback all at once right before the self-assessment ratings may have raised awareness of their performance.

Posttest Procedural Knowledge

Unlike previous experiments, children's percent correct on procedural knowledge at posttest was highest with immediate feedback ($M = 86\%$, $SE = 4\%$), lower with summative feedback ($M = 78\%$, $SE = 5\%$), and lowest with no feedback ($M = 65\%$, $SD = 8\%$). There were significant, positive effects of immediate feedback, $F(1, 67) = 8.20$, $p = .006$, $\eta_p^2 = .11$, and summative feedback, $F(1, 67) = 4.09$, $p = .04$, $\eta_p^2 = .06$, relative to no feedback. A follow-up analysis revealed no significant difference between the two feedback types, $p = .38$.

However, scores on procedural knowledge at posttest were high and not normally distributed. Across conditions, the majority of children (75%) solved 50% correct or higher with a full 47% of children solving all of the items correctly. I used binomial logistic regression to predict the log of the odds of scoring 100% correct. The results are displayed in Figure 5.

There was a significant, positive effect of immediate feedback. Children in the immediate-feedback condition were *more* likely than children in the no-feedback condition to score 100% on the posttest (63% vs. 38%), $\beta = 1.32$, $z = 1.97$, *Wald* $(1, N = 73) = 3.89$, $p = .05$, $OR = 3.74$. There was no effect of summative feedback. Children in the summative-feedback condition were just as likely as children in the no-feedback condition to score 100% on the

posttest (40% vs. 38%), $\beta = 0.34$, $z = 0.52$, *Wald* $(1, N = 73) = 0.28$, $p = .60$, *OR* $= 1.40$. A follow-up analysis revealed no significant differences between the two feedback conditions, $p = .12$. Exploratory analyses revealed that the conclusions remain unchanged after excluding children who received only positive feedback. Further, the pattern of results was relatively consistent across the learning and transfer items, though more pronounced for transfer. Overall, children with knowledge of a correct strategy learned best from immediate feedback. Summative feedback also improved learning, but not mastery, relative to no feedback.

As in previous experiments, the conceptual knowledge results are reported in full in Appendix C. Feedback had no reliable impact on conceptual knowledge.
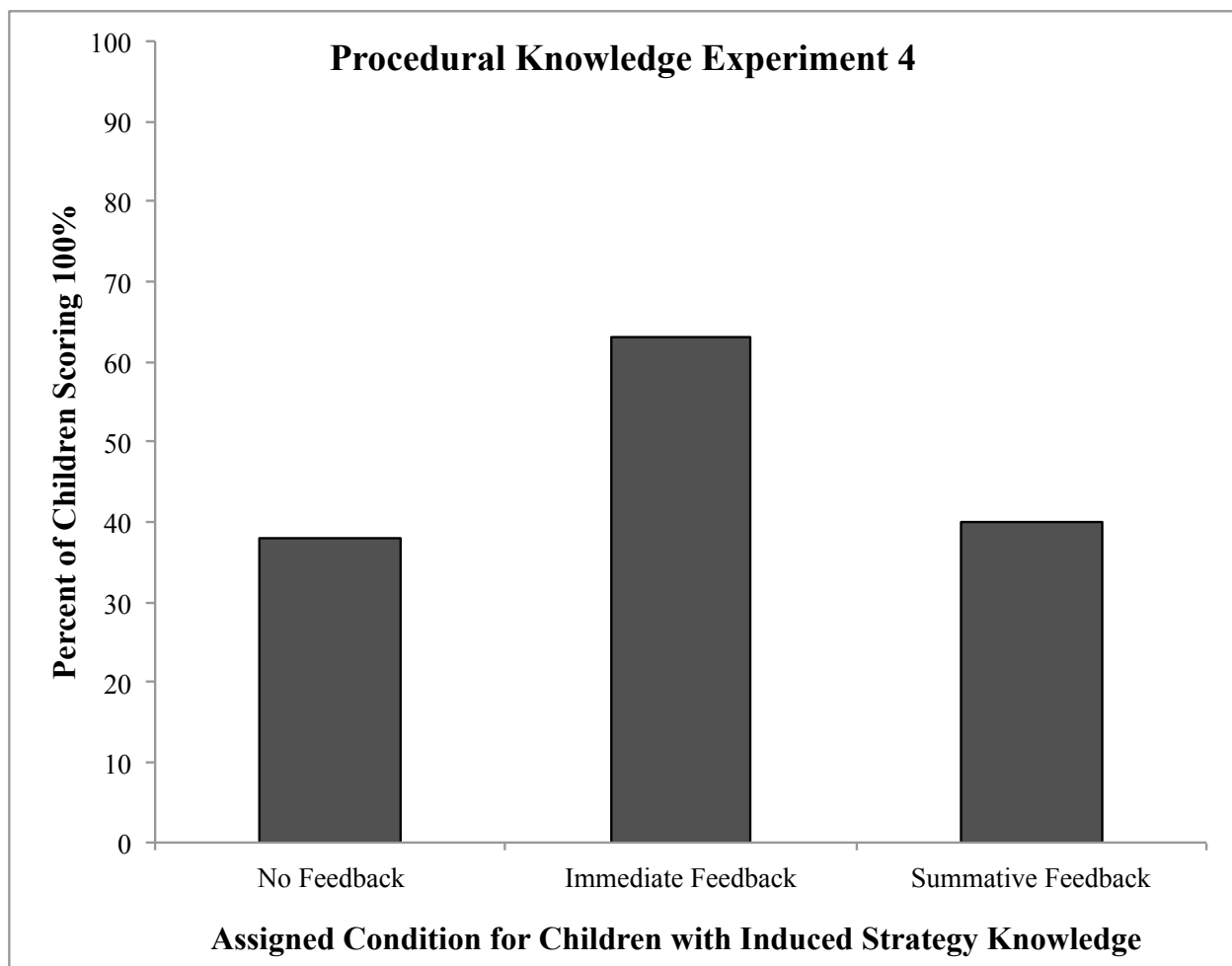
Figure 5. Procedural knowledge at posttest in Experiment 4

In stark contrast to Experiments 1, 2, and 3, the results from this experiment show positive effects of feedback on mathematics learning for children with higher prior knowledge. During the intervention, immediate feedback had neutral effects relative to no feedback on strategy use, cognitive load, and ratings of self-assessment. However, on the next-day posttest, both immediate and summative feedback resulted in higher procedural knowledge than no feedback. The positive effects of immediate feedback were particularly robust, facilitating learning as well as mastery of the material relative to the other conditions.

The results presented here are consistent with a large body of research demonstrating positive effects of feedback on learning and performance relative to a no feedback control (e.g., Bangert-Drowns et al., 1991; Hattie & Timperley, 2007). They also support the notion that computer-generated feedback may be particularly effective (Kluger & DeNisi, 1996).

Most importantly, these results provide some insight into when feedback may have positive versus negative effects. Specifically, the present results suggest that feedback may be more effective in environments that reduce individualized attention on the self. Indeed, Experiments 2 and 4 were nearly identical with a few key exceptions: Experiment 4 removed the presence of the experimenter during problem solving, removed the explicit right/wrong judgment from the feedback message, and delayed the posttest by a day. With these changes made, the results changed drastically from showing significant negative effects of feedback in Experiment 2 to showing significant positive effects of feedback in Experiment 4. These results are consistent with the hypothesis that "feedback interventions that contain cues that direct attention to the self, or that are given in a self-threatening environment, will produce weak or even

negative effects on performance" (Kluger & DeNisi, 1998, p. 69). It should be noted, however, that summative feedback produced significantly lower ratings of positive self-assessment and still resulted in higher learning on the posttest relative to no feedback. Thus, the direct role of ego-threat and self-assessment in changing feedback's impact needs further investigation.

Additionally, it is possible that reduced individualized attention may impact problem solving *without* feedback. That is, the changes in Experiment 4 may not have increased the effectiveness of the feedback conditions, but rather *decreased* the effectiveness of the no-feedback condition. One possibility is that the experimenter's presence and attention provides accountability for these children in the no-feedback condition, which motivates them to think more carefully about the problems and solution strategies. If this is the case, then person-presence and individualized attention may be a double-edged sword. When feedback is provided, the person-presence may focus too much attention on the evaluative aspect of the feedback, create a self-threatening environment, and ultimately hinder learning. However, when no feedback is provided, the person-presence may motivate children to focus attention on the task in a way that they may not do on their own, and ultimately facilitate learning. Future research is needed to experimentally test this speculation regarding the role of person-presence.

CHAPTER VI

GENERAL DISCUSSION

The current study is the first to provide causal evidence that differences in prior knowledge can lead to varying effects of feedback during mathematics problem solving. Specifically, for children with low prior knowledge, feedback had positive effects on learning relative to no feedback. However, for children with moderate prior knowledge, feedback had negative effects on learning relative to no feedback under some conditions.

In Experiment 1, children were randomly assigned to condition based on a crossing of two factors in a one-on-one tutoring session: induced strategy knowledge (yes vs. no) and feedback (present vs. absent). Before problem solving, children in the induced strategy knowledge condition were given instruction on a correct strategy. During problem solving, some children were given immediate, verification feedback on their answers and some were not. Children with no knowledge of a correct strategy exhibited higher procedural knowledge on a posttest if they received feedback than if they did not. In contrast, children with induced strategy knowledge exhibited higher procedural knowledge if they did *not* receive feedback.

In Experiment 2, all children were given instruction on a correct strategy in a one-on-one tutoring session. During problem solving, some children received no feedback, some children received immediate correct-answer feedback after each problem, and some children received summative correct-answer feedback after all the problems had been solved. Consistent with Experiment 1, children exhibited higher procedural knowledge on a posttest if they did *not* receive feedback. There were no differences between immediate- and summative-feedback.

In Experiment 3, all children were given instruction on a correct strategy in small groups within their classrooms. During problem solving, groups received either no feedback, immediate correct-answer feedback, or summative correct-answer feedback. Children in all groups exhibited similar procedural knowledge on a posttest. However, more children exhibited mastery in procedural knowledge on a one-week retention test in the no-feedback condition. Again, there were no differences between immediate- and summative-feedback.

In Experiment 4, all children were given instruction on a correct strategy in a one-on-one tutoring session. During problem solving, children received no feedback, immediate correct-answer feedback, or summative correct-answer feedback from the computer in the absence of an experimenter. In contrast to the previous experiments, children exhibited higher procedural knowledge on a next-day posttest if they received feedback (immediate or summative) than if they did not, and the immediate-feedback condition produced the most children at mastery.

In the discussion that follows, I integrate these results with previous research, offer potential explanations for the mixed effects of feedback, and discuss several limitations that suggest directions for future research on the effects of feedback and prior knowledge.

Pinpointing Prior Knowledge as a Key Moderator

These findings contribute to the feedback literature in several ways. For example, they address a call to explore the impact of feedback in relation to individual differences generally (Hattie & Gan, 2011) and prior knowledge specifically (Mason & Bruning, 2001; Shute, 2008). Here, I used a pre-familiarization technique, in which some children were exposed to a correct strategy and others were not (Petersen & McNeil, 2013; Rey & Buchwald, 2011). This avoids confounding variables, allows for random assignment, and establishes a causal relation. I also

examined a specific type of prior knowledge, knowledge of domain-specific solution strategies, and specified the level of knowledge at which the moderation occurred. I found that only true novices benefitted from feedback during problem solving. Children with moderate knowledge (i.e., used correct and incorrect strategies inconsistently) often benefitted from no feedback.

This finding is consistent with research demonstrating that one instructional method is often not best for all learners (Cronbach & Snow, 1977), and it highlights the need to consider individual differences. Indeed, instructional interventions "are likely to be different for different participants. If individual differences are not examined, average treatment differences may mask and thus miss important special effects" (Snow, 1996, p. 545). For example, expertise reversal effects occur when instructional techniques that are effective for novices lose their benefits for more experienced learners (Kalyuga, 2007). Further, the reversal is often related to levels of instructional guidance, such that low-knowledge learners benefit from strong guidance and support, but higher-knowledge learners benefit from little to no guidance and support (Kalyuga et al., 2003). Experiment 1 provides another example of this phenomena. Specifically, low-knowledge learners benefitted from the provision of verification feedback, but learners with higher prior knowledge actually learned more when no feedback was provided. Further, this reversal occurred well before the children were experts in the domain.

The Positive Effects of Feedback for Low-Knowledge Learners

A key result of this work and my previous work (Fyfe et al., 2012) is the strong, positive effect of verification feedback for children with low prior knowledge. In Experiment 1, simply telling the no-knowledge children that their answers were right or wrong allowed them to go from solving zero problems correctly at pretest to solving half of the problems correctly on the

posttest, both a significant and meaningful increase. Further, the effects of verification feedback were so positive for the no-knowledge group that there were no statistical differences between no-knowledge children and strategy-knowledge children who received feedback.

This result is consistent with previous studies demonstrating substantial benefits on learning and development that can occur from the provision of minimal feedback. For example, 3-year-old children who were told if they were right or wrong on a card sort task successfully switched from sorting by one rule (e.g., color) to sorting by a new rule (e.g., shape), but those who did not receive this feedback continued to perseverate (Bohlmann & Fenson, 2005). Similarly, 6- and 7-year-old children generally struggle with Piagetian conservation tasks, but providing verification feedback improves performance on subsequent trials (Brainerd, 1972). Finally, verification feedback enhances 8- and 9-year-old children's ability to predict the movement of a balance beam relative to no feedback (Tudge et al., 1996). Clearly, even minimal feedback can benefit learning and serve as an impetus for cognitive change.

The positive effects of feedback for the low-knowledge children are particularly insightful given the use of verification feedback alone. Recall, the amount of feedback information can vary. At the most basic level, feedback can verify the correctness of a response (i.e., right/wrong). However, feedback can also include the correct response or a number of additional elaborations (e.g., explanation of why a response is correct; Kulhavy & Stock, 1989). A common assumption is that verification feedback alone is relatively ineffective (e.g., Bangert-Drowns et al., 1991; Butler, Godbole, Marsh, 2013). Indeed, a number of studies have found little to no benefit of verification feedback compared to no feedback and compared to correct-answer feedback (Gilman, 1969; Pashler, Cepeda, Wixted, & Rohrer, 2005; Roper, 1977). However, in other cases (including in the present study), right/wrong verification feedback has

led to substantial learning (Brainerd, 1972; Fyfe et al., 2012; Wentling, 1973). In general, the point is not to suggest that verification feedback is the most effective form of feedback or that it will always benefit learning, but rather to demonstrate that verification feedback can impact learners in a positive way. Clearly, seemingly minor modifications to the input provided during problem solving can substantially benefit the knowledge that learners construct.

Why might verification feedback benefit low-knowledge learners? The intervention results from Experiment 1 help explain why feedback had positive effects for these novices and also inform several theorized functions of feedback. First, verification feedback prevented no-knowledge children from perseverating on the same incorrect strategy, as has been found in previous research (e.g., Fyfe et al., 2012; Bohlmann & Fenson, 2005). That is, feedback encouraged children to entertain alternative approaches to the problems and may have reduced *mindlessness*, which is commitment to a "single, rigid perspective and…oblivious to alternative ways of knowing" (Langer, 2000). Second, verification feedback facilitated children's generation of at least one correct strategy. Indeed, discovering new problem-solving procedures is a key source of cognitive change and can be a strong predictor of subsequent performance (e.g., Rittle-Johnson, 2006; Siegler & Shipley, 1995). However, the findings suggest that feedback may not always play these positive roles once children know a correct strategy.

The Mixed Effects of Feedback for Higher-Knowledge Learners

A second key result of this work is the negative effects of verification and correct-answer feedback for the strategy-knowledge group under some conditions. In Experiments 1, 2, and 3, problem solving *alone* was often more effective for learners with some prior knowledge than problem solving with feedback—even minimally intrusive verification feedback. Importantly,

these learners were not experts in the domain. Thus, the differences cannot be explained by more general expert-novice differences. Rather these learners were briefly exposed to one correct strategy, and many children continued to use a mix of both correct and incorrect strategies during problem solving. Negative effects of feedback have been reported in other studies as well (Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996), with several hypotheses about why they may occur. In contrast to these negative effects, in Experiment 4, feedback had substantial positive effects relative to no feedback for children with higher prior knowledge. Below, I outline several potential mechanisms underlying the potential negative effects for the strategy-knowledge group, and why these effects may have reversed in the final experiment.

First, I have direct evidence that immediate verification and correct-answer feedback increased the use of various incorrect strategies relative to no feedback in Experiments 1 and 2, which is consistent with our prior work (Fyfe et al., 2012). For the no-knowledge group, this negative effect was offset by a positive result, namely the introduction of a correct strategy into their repertoire. For the strategy-knowledge group, the use, strengthening, and potential introduction of different incorrect strategies likely had detrimental effects. Incorrect strategies compete with existing strategies and can reduce the frequency of correct strategy use (Siegler & Shipley, 1995). Thus, on the posttest, these children could select from a number of different strategies, including the correct one they knew, but also a number of different incorrect ones. Further, previous research suggests that, in some cases, frequent shifts in strategy use are negatively related to learning (e.g., Coyle & Bjorklund, 1997; McGilly & Siegler, 1989). Thus, the introduction and/or strengthening of incorrect strategies may help explain why the strategy-knowledge children were hindered by immediate feedback relative to the no-feedback control.

However, the results from Experiments 2, 3, and 4 suggest there are other mechanisms at work. Summative feedback was provided after problem solving and had no reliable impact on strategy generation relative to no feedback; yet, it still resulted in lower learning in Experiments 2 and 3. Thus, there must be other aspects of feedback that negatively impact learners.

One possibility is that feedback results in the processing of redundant information for learners with higher prior knowledge and overloads their cognitive resources. Monitoring and evaluating feedback takes place in working memory, a short-term system that enables individuals to control, maintain, and regulate a limited amount of task-relevant information (Miyake & Shah, 1999). When there are high demands on working memory, the system can overload and hinder learning (Sweller et al., 1998). Some forms of instructional guidance are thought to cause cognitive overload for higher-knowledge learners because the information provided is redundant with their existing knowledge. The redundant information is still processed in working memory, which uses resources that could be devoted to more germane tasks (Sweller et al., 2011). Indeed, the redundancy principle is often used to explain expertise reversal effects (see Kalyuga, 2007). The idea is that low-knowledge learners need instructional guidance to make progress, but once learners gain sufficient knowledge, the guidance becomes redundant and burdensome to process.

In the current study, it seems safe to assume that the feedback was not redundant for the no-knowledge group. However, for the strategy-knowledge group, the verification and correct-answer feedback messages provided some already-known information. For example, these children solved many problems correctly using the instructed strategy. On these trials, it seems likely that the feedback was redundant and detracting from task-relevant processing. Further, the information would be redundant regardless of whether it was provided immediately or after the problem-solving task. Children's ratings of task difficulty fail to provide evidence for this

71

interpretation. Although, cognitive load is a broad construct that can be measured in multiple ways (see Paas et al., 2003), and appears more difficult to assess in children.

The positive effects of feedback in Experiment 4 provide tentative support for the role of the redundancy effect. First, the computer-generated feedback was likely less redundant than the person-generated feedback in Experiments 2 and 3 because it removed the auditory component. In Experiments 1 and 2, the feedback was presented both verbally by the experimenter and visually on the computer screen. One of the recommended principles for reducing cognitive load in multimedia environments is to avoid presenting identical information in print and in speech (Mayer, 2008; Mayer & Moreno, 2003). Thus, if redundancy was the reason that feedback hindered learning, reducing the redundancy should enhance the benefits of feedback. Second, the computer-generated feedback gave learners more control than the person-generated feedback. Specifically, children could process the feedback at their own pace or choose to ignore the information completely. Indeed, at least one study suggests that adult learners benefit from the opportunity to control feedback, as they often seek it when necessary and skip it when not (Hays, Kornell, & Bjork, 2010). Thus, feedback may have had positive effects in Experiment 4 precisely because the presentation method reduced the consequences of redundancy.

A second possibility is that feedback reduces self-confidence in higher-knowledge learners and ultimately hinders learning. Indeed, one of the most comprehensive theories of feedback suggests that negative effects arise when feedback directs attention to the self, as attention on the self can produce affective responses that interfere with learning and performance (Kluger & DeNisi, 1996). For example, learners may become fixated on whether they were right or wrong (and how that reflects on their abilities) rather than on ways to improve.

Learners with some prior knowledge likely have some expectation of performing well (e.g., Kluger & Adler, 1993), but haven't mastered the task. This may lead to a heightened sensitivity to incorrect responses. For example, feedback on incorrect trials may have produced ego threat (i.e., a threat to one's positive self-image), and decreased children's confidence in their use of the taught strategy. This may have led them to revert to old (incorrect) strategies (either during problem solving or on the posttest). Learners with low prior knowledge may be less susceptible to this attention on the self as they can attribute incorrect responses to their lack of knowledge or experience with no threat to their abilities.

A comparison across experiments provides indirect support for the idea that attention on the self plays a role. For example, feedback had negative effects in Experiments 1 and 2, in which children worked one-on-one with an adult and received verbal feedback on their accuracy. The negative effects were less robust in Experiment 3, in which children worked in small groups in the classroom. One could argue that attention on the self was reduced due to the classroom setting where children's accuracy was less noticeable to the adult. However, the increased attention from peers may have offset the reduced attention from the adult and potentially enhanced the role of ego-threat, particularly for students who took more time on tasks relative to the other students in the group. Thus, the role of attention on the self is ambiguous in a classroom setting. The strongest support for the role of attention on the self comes from Experiment 4. The negative effects of feedback reversed in Experiment 4, in which the adult's presence was removed during problem solving and feedback was provided by a computer. Although children's self-assessment responses did not provide evidence that feedback changed children's self-image, there was a restricted range of scores, which may have limited the usefulness of the measure. Also, it only assessed ratings of positive attributes, so it did not directly measure ego-threat.

In sum, three potential mechanisms for the negative effects of feedback in the current study include changes in strategy use, the redundancy effect, and attention on the self. The negative effects of summative feedback rule out strategy change as the sole explanation. Although the analyses of subjective reports failed to support the roles of redundancy and attention on the self, there is some evidence for their relevance and their roles warrant further experimental investigation. Future research should also examine other potential mechanisms not discussed here. For example, feedback is hypothesized to hinder learning when it reduces mindful processing of the task at hand (Bangert-Drowns et al., 1991). If feedback makes a task seem too easy or if it encourages overreliance on the feedback message itself, learners may reduce efforts to construct an appropriate response (Jacobs & Kulkarni, 1966). Other researchers focus more explicitly on the positive effects of practice without feedback. Indeed, several experiments demonstrate the positive changes in strategy use that occur in the absence of input (e.g., Christianson et al., 2012; Nussbaumer et al., 2014).

Limitations, Future Directions, and Conclusion

Despite the positive contributions of the current study, several limitations suggest directions for future research. Perhaps the most important endeavor for future research is to test the generalizability of these results to different tasks, settings, and populations. Across five experiments (current study and Fyfe et al., 2012), I have found that verification and correct-answer feedback can have negative effects relative to no feedback. But, all of this research has occurred with elementary school children learning to solve math equivalence problems from a novel experimenter. Thus, the generalizability of the conclusions may be relatively limited. For example, feedback may function differently for children learning from a supportive, familiar

74

teacher or parent who has established positive norms for dealing with errors. Feedback may also function differently for children in a high-stakes environment, in which they feel more pressure to master the material. Previous research has also found negative effects of feedback, but primarily with adults in non-problem solving domains (see Bangert-Drowns et al., 1991). Experiment 4 provides one boundary condition. Specifically, the negative effects of feedback do not seem to generalize to computer-generated feedback in the absence of a supervising adult. However, more research is needed with various populations, domains, and tasks.

More work is also needed to better understand the negative effects of feedback and the potential contributions of the redundancy effect and attention on the self. For example, one possible way to test the role of attention on the self is to manipulate the source of the feedback provided (e.g., person versus computer) within a single experiment. Another possibility is to manipulate expectations by telling some children that the task is hard and meant for older children. These children should expect to receive negative feedback and not feel threatened by it, in which case they may benefit from it relative to no feedback.

A related issue is to figure out when negative effects appear and how long they last. Findings have been mixed. I have found negative effects of feedback that occur only on an immediate test (Experiment 1), only after a one week delay (Experiment 3), or both immediately and after a delay (Fyfe et al., 2012). One thing to consider is the way in which prior knowledge was operationalized. In the past, I have relied on a pretest to assess learners' prior knowledge of math equivalence. Here, I manipulated learners' prior knowledge in a brief 5- to 10-minute span by providing instruction on a correct problem-solving strategy. It seems likely that these two types of "prior knowledge" vary in certain ways that may influence future performance.

Finally, more work is needed to examine the impact of various feedback types. Together with my prior work (Fyfe et al., 2012), I have found negative effects of feedback for higher knowledge learners using feedback that varies in content (i.e., focused on answers vs. focused on strategies), amount (i.e., verification vs. verification plus the correct answer), and timing (immediate vs. summative). However, these and other features of feedback require further investigation. For example, the *source* of feedback may play a key role in determining whether feedback helps or hinders learning. A comparison of Experiments 2 and 4 suggests that feedback from a computer may be more effective than feedback from a novel person. Another possibility is that feedback from a familiar person may be more effective than feedback from a novel person. Familiar teachers and tutors can form supportive relationships with their students and establish classroom norms that may alter the way feedback is given and received. Specifically, students may not feel threatened receiving negative feedback from a familiar teacher, whom they can ask for help in working through their mistakes.

Other features of feedback may matter as well. For example, the *amount* of information in the feedback message may impact its efficacy. All versions of our feedback have provided relatively minimal input. Other types of feedback provide elaborated information, such as a conceptual rationale of the correct answer or a hint on a correct problem-solving strategy. For example, Narciss and Huth (2006) have found positive effects of "bug-related feedback" in a subtraction task, which flags the specific error made, provides a hint on the correct strategy, and (if errors persist) presents a step-by-step solution strategy along with the correct answer. Elaborated feedback may be more beneficial as it provides more information for error-correction and understanding. On the other hand, elaborated feedback may be more harmful because the larger amount of information requires more cognitive processing, which may result in increased

cognitive interference. I am currently in the process of investigating the impact of more elaborated feedback on middle-school students' algebra problem-solving. The feedback occurs on homework assignments in a computer tutor environment and contains either just the correct answer or the correct answer and an explanation as to why it is correct.

In conclusion, the present study provides causal evidence for a moderator that can help explain both positive and negative effects of some types of feedback under some conditions. Specifically, children with no knowledge of a correct strategy benefitted from verification feedback during problem solving. In contrast, children with induced knowledge of a correct strategy often learned more from problem solving alone, particularly when in the presence of a novel adult tutor. The latter result is consistent with recent research on the need for "productive struggle" during learning (Kapur, 2012; Schwartz et al., 2011). The idea is that students benefit from periods of exploration during which they engage with relevant problems with minimal external guidance. This need for productive struggle has also been recognized by the National Council of Teachers of Mathematics (2014). Too often, "teachers jump in to rescue students by breaking down the task and guiding students step by step through the difficulties. Although well-intentioned, such rescuing undermines the efforts of students, lowers the cognitive demand of the task, and deprives students of opportunities to engage fully in making sense of the mathematics" (NCTM, 2014, p. 48). When people (e.g., teachers, parents, tutors) provide feedback to higher-knowledge learners, it may be another form of "rescuing" that can hinder learning.

Screening Measure
From McNeil, Fyfe, Dunwiddie, & Brletic-Shipley, 2011

Section 1: Equation-Solving

Directions: Try your best to solve each problem. Write the number that goes in the blank.
1. $1 + 5 = \underline{} + 2$
2. $7 + 2 + 4 = \underline{} + 4$
3. $2 + 7 = 6 + \underline{}$
4. $3 + 5 + 6 = 3 + \underline{}$

Scoring: Received one point for each problem solved using a correct strategy. Solutions within plus/minus one of the correct answer were coded as a correct strategy.

Section 2: Equation-Encoding

Directions: I'll show you a problem for five seconds, and I want you to try to remember the problem. After I hide the problem, I want you to write exactly what you saw.
1. $7 + 1 = \underline{} + 6$
2. $3 + 5 + 4 = \underline{} + 4$
3. $4 + 5 = 3 + \underline{}$
4. $2 + 3 + 6 = 2 + \underline{}$

Scoring: Received one point for each problem accurately reconstructed. All numerals, operators, and equal sign had to be in proper location.

Section 3: Equal Sign Definition

Directions: The following questions are about this math symbol (=). Write your answers to the questions below.

1. What is the name of the symbol?
2. What does this math symbol mean?
3. Can it mean anything else?

Scoring: Received one point for providing a relational definition of the equal sign, such as "the same amount" or "two sides are the same."

Mathematical Equivalence Knowledge Measure
Based on Matthews, Rittle-Johnson, McEldoon, & Taylor, 2012 and Rittle-Johnson,
Matthews, Taylor, & McEldoon, 2011

Procedural Knowledge Scale

Directions: Try your best to solve each problem. Write the number that goes in the box.

1. $8 = 6 + \underline{}$
2. $3 + 4 = \underline{} + 5$
3. $3 + 7 + 6 = \underline{} + 6$
4. $7 + 6 + 4 = 7 + \underline{}$
5. $\underline{} + 2 = 6 + 4$
6. $8 + \underline{} = 8 + 6 + 4$
7. $5 + 6 - 3 = 5 + \underline{}$
8. $5 - 2 + 4 = \underline{} + 4$

Scoring: Received one point for each problem solved using a correct strategy. Solutions within plus/minus one of the correct answer were coded as a correct strategy.

Conceptual Knowledge Scale

Directions: I'll show you a problem for five seconds, and I want you to try to remember the problem. After I hide the problem, I want you to write exactly what you saw.

1. $4 + 3 + 9 = 4 + \underline{}$
2. $8 + 6 + 3 = \underline{} + 2$

Scoring: Received one point for each problem accurately reconstructed. All numerals, operators, and equal sign had to be in proper location.

Directions: Write your answers to the following questions.

3. What does the equal sign (=) mean? Can it mean anything else?

Scoring: Received one point for providing a relational definition of the equal sign, such as "the same amount" or "two sides are the same."

4. Which answer choice below would you put in the empty box to show that ten cents is the same amount of money as one dime? 10 cents   One dime
(a) 10¢   (b) =   (c) +   (d) don't know

Scoring: Received one point for selecting (b) =.

Directions: For each example, decide if the number sentence is true. In other words, does it make sense? After each problem, circle True, False, or Don't Know.
5a. 3 = 3
5b. 5 + 3 = 8
5c. 31 + 16 = 16 + 31
5d. 7 + 6 = 6 + 6 + 1
5e. 5 + 5 = 5 + 6
5f. 7 = 3 + 4

Scoring: Received one point for circling true for both 5a and 5f. Received one point for circling true for both 5c and 5d.

Directions: Is this a good definition of the equal sign? After each definition, circle Good, Not Good, or Don't Know.

6a. The equal sign means two amounts are the same.
6b. The equal sign means add.
6c. The equal sign means the answer to the problem.

Scoring: Received one point for circling Good for 6a.

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

Scoring: Received one point for selecting 6a from the previous question.

8. Decide if 6 + 4 = 5 + 5 is true or false. Then explain how you know.

Scoring: Received one point for circling true and for providing a relational explanation, such as "both sides have 10" or "if you take one from 6 and add it to 4, you get a 5 and a 5."

9. What does the equal sign mean in this statement: 1 dollar = 100 pennies?

Scoring: Received one point for providing a relational definition of the equal sign, such as "the same amount" or "two sides are the same."

Appendix C

Conceptual Knowledge Results

Experiment 1: Conceptual Knowledge
In addition to the procedural knowledge scale at posttest, I also administered a 10-item conceptual knowledge scale (α = .73). It assessed two key concepts of math equivalence: the meaning of the equal sign and the structure of equations. See Appendix B for a list of the items and scoring criteria. We coded the items requiring a written explanation (e.g., equal sign definition). A second rater coded 30% of the responses and interrater agreement was high (kappas = .94 – .98). An identical conceptual knowledge measure was also administered at the delayed retention test. To evaluate children's performance on the conceptual knowledge measure at posttest and retention test, I conducted two separate ANCOVAs with strategy knowledge (yes vs. no) and feedback (present vs. absent) as between-subject variables. Children's age and screening measure scores were included as covariates. Scores are reported in Table A1.

Table A1. Performance on conceptual knowledge by condition in Exp. 1

Time	Knowledge Condition	No Feedback	Feedback
Posttest	No Knowledge	32 (5)	39 (5)
	Strategy Knowledge	43 (5)	43 (4)
Retention	No Knowledge	35 (5)	47 (5)
	Strategy Knowledge	48 (5)	45 (5)

Note. Scores are estimated marginal means. Standard errors are in parentheses.

At the posttest, there was no main effect of feedback, $F(1, 93) = 0.46$, $p = .50$, $\eta_p^2 = .01$. There was a marginal effect of knowledge, $F(1, 93) = 2.86$, $p = .09$, $\eta_p^2 = .03$, with children in the strategy knowledge condition exhibiting higher conceptual knowledge ($M = 43\%$, $SE = 3\%$) than children in the no-knowledge condition ($M = 35\%$, $SE = 3\%$). However, there was no feedback by knowledge interaction, $F(1, 93) = 0.60$, $p = .44$, $\eta_p^2 = .01$. At the retention test, there were no main effect of feedback, $F(1, 90) = 0.92$, $p = .34$, $\eta_p^2 = .01$, or knowledge, $F(1, 90) = 1.33$, $p = .25$, $\eta_p^2 = .02$. There was also no feedback by knowledge interaction, $F(1, 90) = 2.28$, $p = .14$, $\eta_p^2 = .03$. In general, feedback had little impact on children's conceptual knowledge.

Experiment 2: Conceptual Knowledge
I also administered the 10-item conceptual knowledge measure in Experiment 2 (α = .69). It was identical to the measure in Experiment 1. We coded the items requiring a written explanation (e.g., equal sign definition). A second rater coded 30% of the responses and interrater agreement was high (kappas = .94 – .98). To evaluate children's performance on the conceptual knowledge measure at posttest, I ran an ANCOVA with condition as a between-subject variable. Specifically, condition was dummy coded with immediate feedback and summative feedback entered into the model, and no feedback as the reference group. Children's age and screening measure scores were included as covariates.

At the posttest, children in the no-feedback condition exhibited higher conceptual knowledge ($M = 53\%$, $SE = 4\%$) than children in the immediate-feedback ($M = 41\%$, $SE = 4\%$) and summative-

feedback ($M = 40\%$, $SE = 4\%$) conditions. Indeed, there were significant negative effects of immediate-feedback, $F(1, 84) = 4.80$, $p = .03$, $\eta_p^2 = .05$, and summative-feedback, $F(1, 84) = 5.03$, $p = .03$, $\eta_p^2 = .06$, relative to no feedback. In general, feedback had negative effects on conceptual knowledge relative to no feedback for children with some prior knowledge.

Experiment 3: Conceptual Knowledge
In Experiment 3, I administered a brief 5-item conceptual knowledge measure ($\alpha = .56$). It included four encoding items and one define-the-equal-sign item. I used a brief measure relative to Experiments 1 and 2 to reduce the amount of time spent testing. We coded the encoding items for accurate reconstructions as well as the equal sign item for relational definitions. A second rater coded 30% of the responses and interrater agreement was high (kappas = .86 – .97). An identical measure was administered at the retention test. To evaluate children's performance on the conceptual knowledge measure at posttest and retention test, I used multi-level modeling to account for the variance due to small group. The model had two levels: (1) the individual level and (2) the small-group level. Condition was dummy coded with immediate feedback and summative feedback entered into the models, and no feedback as the reference group. Screening measure scores and age were included as covariates.

At posttest, conceptual knowledge similar across conditions, but somewhat higher in the summative-feedback condition ($M = 61\%$, $SE = 3\%$), than in the immediate-feedback condition ($M = 56\%$, $SE = 3\%$) and the no-feedback condition ($M = 53\%$, $SE = 3\%$). There was a marginal, positive effect of summative-feedback relative to no feedback, $\beta = 7.03$, $p = .07$, but no effect of immediate feedback, $\beta = 1.93$, $p = .60$. A follow-up analysis revealed no difference between the two feedback types, $\beta = –5.10$, $p = .18$. At retention, conceptual knowledge was similar in the summative- ($M = 64\%$, $SE = 3\%$), immediate- ($M = 63\%$, $SE = 3\%$), and no-feedback ($M = 60\%$, $SE = 3\%$) conditions. There was no effect of summative-feedback, $\beta = 2.80$, $p = .39$, and no effect of immediate-feedback, $\beta = 1.84$, $p = .57$. A follow-up analysis revealed no difference between the two feedback types, $\beta = –0.96$, $p = .76$.

Experiment 4: Conceptual Knowledge
In Experiment 4, I administered the 10-item conceptual knowledge measure ($\alpha = .72$). It was identical to the measure in Experiments 1 and 2. We coded the items requiring a written explanation (e.g., equal sign definition). A second rater coded 30% of the responses and interrater agreement was high (kappas = .95 – .96). To evaluate children's performance on the conceptual knowledge measure at posttest, I ran an ANCOVA with condition as a between-subject variable. Specifically, condition was dummy coded with immediate feedback and summative feedback entered into the model, and no feedback as the reference group. Children's age and screening measure scores were included as covariates.

At posttest, conceptual knowledge was similar in the immediate- ($M = 48\%$, $SE = 6\%$), summative- ($M = 56\%$, $SE = 5\%$), and no-feedback ($M = 55\%$, $SE = 4\%$) conditions. There was no effect of immediate-feedback, $F(1, 67) = 1.00$, $p = .32$, $\eta_p^2 = .02$, and no effect of summative-feedback, $F(1, 67) = 0.39$, $p = .53$, $\eta_p^2 = .01$, relative to no feedback. A follow-up analysis revealed no difference between the two feedback types, $F(1, 67) = 2.64$, $p = .11$, $\eta_p^2 = .04$.

Cognitive Load Measure and Validation

Rationale

Cognitive load is rarely measured in young children and I am not aware of any validated scales for use with elementary-school children. I developed a 9-item measure to be suitable for young children. These items were developed based on the theoretical construct and definition of cognitive load, the language limitations of young children, and pilot work with a group of elementary-school children. The measure included three subscales intended to tap distinct aspects of cognitive load: mental effort, mental frustration, and task difficulty. Each of the three subscales were adapted from existing items used with older children or adults. The mental effort and mental frustration scales were based on modified items from the NASA Task Load Index (Hart & Staveland, 1988, see also Fyfe, DeCaro, & Rittle-Johnson, 2015), a measure used in previous studies with adults to assess cognitive load (Rey & Buchwald, 2011). The task difficulty scale was based on a common subjective rating scale used to assess cognitive load with adolescents and adults (see Paas, Tuovinen, Tabbers, & Van Gerven, 2003). I checked the validity of the measure in several ways. I examined its similarity with a common rating scale used with adults and whether it predicted learning outcomes. I also measured its association with relevant variables. For example, I expected cognitive load to negatively relate to motivation, as theory suggests that tasks that produce cognitive overload can reduce motivation. I also expected cognitive load to have no relation to other cognitive variables that assess different aspects of cognitive processing (e.g., working memory, retrieval fluency). Results are outlined below.

Participants

The results below are based on the final sample in Experiment 1, which contained 108 second- and third-grade children (*M* age = 8.4 yrs, *min* = 7.2 yrs, *max* = 9.8 yrs; 67 girls, 41 boys).

Measures

Cognitive Load. Children's cognitive load was assessed using a 9-item measure I have developed to be suitable for young children (see Table A1). The measure includes three three-item subscales intended to tap distinct aspects of cognitive load: mental effort, mental frustration, and task difficulty. Children responded to each item by circling their answer on a 4-point scale: *strongly disagree, disagree, agree, strongly agree*. This response scale has been used in previous research with elementary-school children (Frantom, Green, & Hoffman, 2002). For each item, children's response was assigned a number from one to four and scores were formed by averaging their responses across the three subscale items. A tenth item was administered for validation purposes. It was adopted from prior adult studies in the cognitive load literature (see Paas et al., 2003). It read: "How easy or difficult was this math task to understand?" Children responded on a 7-point scale ranging from *extremely easy* to *extremely difficult*. Children were assigned a score from one to seven. These items were administered during a one-on-one tutoring session immediately following a mathematics problem-solving activity in the same fixed order.

Table D1. Cognitive load items.

Construct Tapped	Item
Mental Effort	1. I had to think hard to do this math work.
	2. I had to keep track of a lot of things at once to do this math work.
	3. I had to think about a lot of things to do this math work.
Frustration	1. I was stressed and irritated when I did this math work.
	2. When I did this math work I felt calm and relaxed.
	3. I was discouraged and annoyed when I did this math work.
Task Difficulty	*1. This math work was presented in an easy way to understand.*
	2. Compared to other math work I've done, this math work was hard.
	3. This math work was very confusing.

Note. Italicized items are reverse scored.

Motivation. We administered three items from the interest and enjoyment scale of the Intrinsic Motivation Inventory (Ryan, 1982). The items were as follow: "I enjoyed solving the math problems very much." "These math problems were fun to do." "The math problems were very interesting." Children responded to each item by circling their answer on a 4-point scale from *strongly disagree* to *strongly agree*. For each item, children's response was assigned a number from one to four and motivation scores were formed by averaging their responses across the three items. Motivation was assessed immediately following the cognitive load items.

Working Memory Capacity. We measured working memory capacity, which supports learners' ability to actively select, regulate, and process task-relevant information, using the backward digit-span task (Wechsler, 2003). Children were read a series of numbers at a rate of one per second and were asked to repeat the numbers in reverse order. Number series length began at two and ended at a maximum of eight. There were two items per series length. The task was discontinued when a child recalled both items in a series of a given length incorrectly. Children received one point for each series recalled correctly in backward order. Working memory was assessed immediately after the posttest.

Retrieval Fluency. We also measured retrieval fluency (Gaddes & Crocket, 1975)—the controlled search and retrieval of information from long-term memory. Children were asked to name as many items from a category (i.e., "animals" and "things to eat") as possible within a one-minute span. Children received one point for each distinct item named in a category. Scores from each category were averaged together to form a fluency score. Fluency was assessed immediately following the working memory capacity task.

Results
Descriptive Statistics. Table A2 contains the descriptive statistics for the cognitive load items. The task difficulty scale had desirable qualities. The mean was near the middle of the response scale, the scores were sufficiently variable, and the distribution was relatively symmetric with a skewness value close to zero. Further, each individual item was related to the total scale score as indicated by moderate item-total correlations. Neither the mental effort or frustration scale exhibited these positive qualities to the same degree. I also explored an aggregate scale in which we used all nine items.

Table D2. Descriptive statistics for cognitive load items

Item	Item-Total Correlation	Alpha	Mean (out of 4)	SD	Skewness
Mental Effort Scale	--	.33	3.07	0.58	-0.41
Item 1	.13	--	3.19	0.86	-0.84
Item 2	.14	--	2.94	0.93	-0.58
Item 3	.29	--	3.07	0.88	-0.73
Frustration Scale	--	.58	1.89	0.68	0.63
Item 1	.50	--	2.06	1.01	0.49
Item 2	.32	--	1.97	0.87	0.58
Item 3	.37	--	1.65	0.89	1.41
Task Difficulty Scale	--	.77	2.29	0.85	0.24
Item 1	.56	--	2.18	0.99	0.39
Item 2	.61	--	2.44	1.04	0.07
Item 3	.63	--	2.25	1.04	0.29
Aggregate Scale	--	.72	2.42	0.53	0.06
Item 1	.26	--	3.19	0.86	-0.84
Item 2	.08	--	2.94	0.93	-0.58
Item 3	.23	--	3.07	0.88	-0.73
Item 4	.55	--	2.06	1.01	0.49
Item 5	.33	--	1.97	0.87	0.58
Item 6	.34	--	1.65	0.89	1.41
Item 7	.51	--	2.18	0.99	0.39
Item 8	.60	--	2.44	1.04	0.07
Item 9	.62	--	2.25	1.04	0.29

Evidence for Reliability. Internal consistency, as assessed by Cronbach's alpha, was high for the task difficulty scale ($\alpha = .77$), somewhat lower for the aggregate scale ($\alpha = .72$), and low for the mental effort scale ($\alpha = .33$) and frustration scale ($\alpha = .58$). Further, the task difficulty items were also all positively and significantly correlated with each other: items 1 and 2, $r(106) = .48$, $p < .001$, items 1 and 3, $r(106) = .51$, $p < .001$, items 2 and 3, $r(106) = .57$, $p < .001$. For the mental effort scale, the inter-item correlations were .00, .20 and .21, only two of which were statistically significant. For the frustration scale, the correlations were .17, .35, and .42, only two of which were statistically significant.

Evidence for Validity. I examined the relation between children's subjective ratings on the three cognitive load scales (and the aggregate scale) with their ratings on an existing measure of cognitive load. I also examined the relation between these ratings and other relevant variables. Correlations are shown in Table A3. As shown in the table, ratings on the task difficulty scale had a strong, positive correlation with ratings on the existing cognitive load item (Adult CL) demonstrating good convergent validity. Ratings on the frustration scale were also correlated with the existing measure, but ratings on the effort scale were not. Further, the task difficulty scale and the Adult CL item were negatively correlated with scores on the posttest, supporting the idea that children who found the task more difficult during the intervention indeed knew less at posttest. The task difficulty scale and the Adult CL item were also negative correlated with intrinsic motivation scores, supporting the idea that children who found the task more difficult

were also less motivated by the task during the intervention. Importantly, task difficulty scores were unrelated to different cognitive constructs (i.e., working memory capacity and retrieval fluency) demonstrating some discriminant validity.

Summary. Overall, several pieces of evidence support the reliability and validity of the task difficulty scale for assessing children's subjective cognitive load. However, the same was not true for the mental effort scale and the frustration scale. These latter scales were dropped.

Table D3. Correlations between cognitive load scales and relevant variables.

	Adult CL Item	Child's Age	Intrinsic Motivation	WM Capacity	Retrieval Fluency	Posttest Score
Adult CL Item	--	.20*	-.46**	-.12	.03	-.28**
Mental Effort	.02	.03	.15	-.07	.03	-.04
Frustration	.39**	.08	-.37**	-.07	-.06	-.18
Task Difficulty	.64**	-.05	-.31**	-.13	-.12	-.23*
Aggregate Scale	.52**	.02	-.27**	-.13	-.08	-.22*

Note. Adult CL Item is the cognitive load item adopted from prior adult studies. Posttest score is children's total score on an assessment of math equivalence understanding. $^*p < .05$, $^{**}p < .01$

Appendix E

Scripts from Experiment 1 Intervention

INTRODUCTION
Today you're going to try to solve some math problems. You will learn a lot of new things, but it won't be easy. You will probably make mistakes. That's okay. The most important thing will be for you to think about the problems and try to understand them. This will give you a chance to practice and improve your abilities in math. These problems are important because if you try your best to understand the problems, you will learn more about math! So what we want you to do is learn new things.

*Proceed to one of two sections. If in "Strategy Knowledge" condition, proceed to procedural instruction. If in "No Strategy Knowledge" condition, proceed to filler instruction.

PROCEDURAL INSTRUCTION (***STRATEGY KNOWLEDGE CONDITION ONLY***)
Now, we're going to go through a short lesson about how to solve these math problems. The problems will look something like this: (Mouse click)

3 + 4 + 2 = 3 + __
Just like this problem, all of the problems we'll work on now will have a blank and we need to figure out what number goes in the blank. There is more than one way to solve this type of problem, but I'm going to show you one way to solve them today.

First, you should see that there are two sides to this problem, one on this side of the equal sign and one on the other side of the equal sign. One side is 3 + 4 + 2 (sweep side) **and the other side is 3 + blank** (sweep side)**.**

Here is how you can solve this problem.
　　　First, you add up the numbers on this side of the equal sign (sweep side)**. Then, find the number that goes in the blank that will make this side** (sweep side) **add up to that same number. Let me say that again. First, you add up all the numbers on this side** (sweep side)**. Then, you find the number to go in the blank that will make this side** (sweep side) **add up to the same number.**
So for this problem, what is 3 + 4 + 2? (Wait for response.)
　　　Right, this side has 9.
And what number goes in the blank to make this side have 9 also? (Wait for response.)
　　　Right, 6, because 3 + 6 is also 9. So our answer is 6. (Type in 6.)

Let's look at another example to make things more clear. (ENTER)

8 + 4 + 7 = 8 + __

Every time you see an addition problem with a blank like this, you can:
　　　First, add up the numbers on the one side of the equal sign (sweep side)**.**

Then, find a number to go in the blank that will make this side (sweep side) **add up to the same number you got on the first side** (sweep side)**.**
So for this problem, what is 8 + 4 + 7? (Wait for response.)
        Right, 19.
And what number goes in the blank to make this side have 19 also? (Wait for response).
        Right, 11, because 8 + 11 is also 19. So our answer is 11. (Type in 11.)

Okay, let's try another example. (ENTER)

5 + 4 + 3 = __ + 5

Can you tell me which side to start this problem on? (Wait for response.)
        We start on the first side with 5 + 4 + 3 (sweep side)**. So, we can add up the numbers on this side** (sweep side)**. Then, find a number to go in the blank that will make this side** (sweep side) **add up to the same number.**
If we add up this side, what do we get? (Wait for response.)
        We get 12.
What is our next step? (Wait for response.)
        To finish, we find a number to go in the blank that will make this side the same number. So we need this side to have 12 altogether. What plus 5 is 12? 7 + 5 = 12. So our answer is 7. (Type in 7.)

Now I want you to think about everything we have talked about so far, and let's look at one more problem together, okay? (ENTER)

2 + 1 + 4 = __ + 4

Can you point to the numbers we're going to start with? (Wait for response.)
        We start on the first side with 2 + 1 + 4 (sweep side)**.**
How should we solve this problem? (Wait for response.)
        We start by adding up the numbers on the first side.
What do the numbers on the first side add up to? (Wait for response.)
        Right, so we know 2 + 1 + 4 = 7.
Now we need to have this side equal 7 also. How many are already on this side (sweep side)?
(Wait for response.)
        This side already had 4. So what plus 4 is equal to 7? 3 + 4 is equal to 7, so 3 is the number that goes in the blank. (Type in 3). (ENTER)

*Proceed to manipulation check

*FILLER INSTRUCTION (***NO STRATEGY KNOWLEDGE CONDITION ONLY***)*
Now, we're going to go through a short lesson together about how to complete a short math activity. The problems will look something like this: (Mouse click)

3 + 4   9

Just like this problem, the problems we'll work on right now will have two boxes. There will be one yellow box and one blue box. Today I'm going to show you how to complete this short math activity.

First, you should see that one of the boxes will have a single number. The other box will have two numbers with a plus sign. For this problem, the yellow box has a 3 + 4 (point) **and the blue box has a 9** (point)**.**

Here is how you can complete this activity.
> **First, you add up the numbers in the box with two numbers and the plus sign** (point)
> **and think about the number they add up to.**
> **Then, decide which box has the bigger number.**

So for this problem, what is 3 + 4? (Wait for response.)
> **Right, the yellow box has 7.**

And what number is in the blue box? (Wait for response.)
> **Right, 9.**

So if this box has 3 + 4 and this box has 9, which box has the bigger number?
> **Right, the blue box has the bigger number.**

Let's look at another one. (Mouse click)

23   19 + 8

Every time you see the two boxes like this, you can:
> **First, add up the numbers in the box with the two numbers and the plus sign** (point).
> **Then, decide which box has the bigger number.**

So for this problem, what is 19 + 8? (Wait for response.)
> **Right, the blue box has 27.**

And what number is in the yellow box? (Wait for response).
> **Right, 23.**

So if this box has 23 and this box has 19 + 8, which box has the bigger number?
> **Right, the blue box has the bigger number.**

Okay, let's try another example. (Mouse click)

9 + 8   14

Can you tell me which box to start with on this problem? (Wait for response.)
> **Right, we start with the box that has two numbers and a plus sign. So we can add up the two numbers in this box** (point) **and then decide whether the yellow box or the blue box has the bigger number.**

If we add up the numbers in the yellow box, what do we get? (Wait for response.)
> **We get 17.**

What is our next step? (Wait for response.)
> **To finish, we decide which box has the bigger number.**

So if this box has 9 + 8 and this box has 14, which box has the bigger number?

Right, the yellow box has the bigger number.

Now I want you to think about what we have done so far, and let's look at one more example together, okay? (Mouse click)

7 + 4   10

Can you point to the box we're going to start with? (Wait for response.)
   We start with the yellow box (point)**.**
How should we complete this problem? (Wait for response.)
   We start by adding up the two numbers in the yellow box.
What do the two numbers add up to? (Wait for response.)
   Right, the yellow box has 11.
And what number is in the blue box (point)**?** (Wait for response.)
   The blue box has 11. So if this box has 7 + 4 and this box has 10, which box has the bigger number?
   Right, the yellow box has the bigger number.

*Proceed to manipulation check.

<u>MANIPULATION CHECK</u>
Turn on digital voice recorder.
Okay. Thanks for listening so carefully. Now, can you use the strategy I just taught you to solve a problem on your own? (Mouse click)

Find the number that goes in the blank to make this number sentence true.

7 + 6 + 2 = 7 + __
(Have child type in response with number pad and then hit ENTER.)
How did you solve that problem?

***If in "No Strategy Knowledge" Condition:
Thanks for all your hard work! (No feedback. Proceed to exploratory problem solving.)

***If in "Strategy Knowledge" Condition:
If CORRECT answer and CORRECT strategy: **Good job! 8 is the correct answer. You used a correct strategy. You** (repeat child's correct strategy)**. Great work!** (Proceed to Exploratory Problem Solving.)

If INCORRECT answer and CORRECT strategy: **Good try, but that is not the correct answer. It sounds like you used a correct strategy. You** (repeat child's strategy)**. But then you made a small arithmetic mistake and got the incorrect answer. Try to solve that problem one more time.** (Continue on to Exploratory Problem Solving.)

If INCORRECT strategy: **Good try, but that is not the correct answer. Remember, to solve this problem correctly, you add up the numbers on the first side of the equal sign. Then,**

you figure out what number goes in the blank to make this side have that same number. Try again to use that correct strategy to solve a problem on your own.

Computer program will automatically proceed to exploratory problem solving OR to an extra manipulation check problem based on child's answer. Then it will go to a READY screen.

Extra Problems:
4 + 5 + 8 = __ + 8
3 + 6 + 5 = 3 + __
9 + 1 + 2 = __ + 2
7 + 3 + 4 = 7 + __

EXPLORATORY PROBLEM SOLVING
Now I'm going to have you practice solving some problems on your own. These are very similar to the problems we just worked on together. For these problems, you need to figure out the number that goes in the box to make the number sentence true. Some of them may seem difficult or unfamiliar. That's okay. Just try your best.

After each problem, I would like you to tell me when you are finished.

Let's look at the first problem. (Mouse click)
Try to figure out the number that goes in the box to make this number sentence true. Here is some scratch paper to use if you want to. When you have your answer, you can type it using this (hand them number pad), **and then press ENTER.**

1)   10 = 3 + __

Can you tell me how you solved that problem?
Record child's strategy on the strategy record sheet. If it is ambiguous, give additional prompt: **I'm not sure I understand. Can you point to the exact numbers that you added or subtracted or tell me the numbers?** Mark any strategies on the record sheet that need confirmed.

***If in "No Feedback" Condition:
Okay, please try the next problem. Do not give feedback. Try not to say anything positive or negative or make any facial expressions that indicate a correct or incorrect answer. Repeat for all twelve problems. Then proceed to subjective questions.

***If in "Feedback" Condition:
If CORRECT answer: (Mouse click) **Good job! You got the right answer. 7 is the correct answer. Let's try another one.**
If INCORRECT answer: (Mouse click) **Good try, but you did not get the right answer. [Child's answer] is not the correct number. Let's try another one.**
Repeat for all twelve problems. Then proceed to subjective questions.

Repeat for remaining problems.
2)   3 + 7 = 3 + __
3)   3 + 7 = __ + 6
4)   3 + 6 = 3 + __
5)   3 + 4 + 8 = __ + 8
6)   5 + 3 + 9 = 5 + __
7)   9 = 3 + __
8)   9 + 7 + 6 = __ + 6
9)   3 + 7 + 8 = __ + 8
10)   7 = 6 + __
11)   4 + 5 + 3 = 4 + __
12)   8 + 3 + 7 = __ + 7

After the last problem, a Subjective Questions screen will appear.

SUBJECTIVE QUESTIONS
Turn off digital voice recorder.
Thanks for all your hard work! I'm interested in what you think about the problems you just solved. There are a few statements that I'll read through with you.

On each one, circle the answer that shows how much you disagree or agree with the sentence. Strongly Disagree means that "NO" you disagree a lot with what the sentence says. Disagree means that NO, you disagree with the sentence, but not a lot. Agree means, YES you agree with the sentence, but not a lot. Strongly agree means YES you agree a lot with what the sentence says.

When solving the problems just now:

1.   **I had to think hard to do this math work. Do you strongly disagree, disagree, agree or strongly agree? Circle the answer that matches how you think.**
2.   **I was stressed and irritated when I did this math work.**
3.   **This math work was presented in an easy way to understand.**
4.   **I had to keep track of a lot of things at once to do this math work.**
5.   **Compared to other math work I've done, this math work was hard.**
6.   **When I did this math work, I felt calm and relaxed.**
7.   **This math work was very confusing.**
8.   **I had to think about a lot of things to do this math work.**
9.   **I was discouraged and annoyed when I did this math work.**

For this item, you're going to circle a different kind of response.
10.   **How easy or difficult was this math task to understand? Was it extremely easy, very easy, easy, not easy or difficult, difficult, very difficult, or extremely difficult?**

Now I'd like to ask you about a few more thoughts and feelings you had while solving the problems on the computer just now. I will tell you some thoughts or feelings that kids

sometimes have. For each one, circle the answer that matches your response. When I was solving problems on the computer just now:

1. **I enjoyed solving the math problems very much.**
2. **These math problems were fun to do.**
3. **The math problems were very interesting.**

END OF INTERVENTION
Give students a short break of about five minutes. Have them pick out a piece of candy. Let student stretch or get a drink of water if necessary. Have them tell you a story, etc.
(There will be a white blank screen followed by a READY screen.) Proceed to posttest.

REFERENCES

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1-18. doi:10.1037/a0021017

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology*, *35*, 127-45. doi:10.1037/0012-1649.35.1.127

Alibali, M. W., Phillips, K. M. O., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development, 24*, 89–101. doi:10.1016/j.cogdev.2008.12.005

Anderson, D. I., Magill, R. A., & Seklya, H. (2001). Motor learning as a function of KR schedule and characteristics of task-intrinsic feedback. *Journal of Motor Behavior, 33*, 59-66.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*, 111-127

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213. doi:10.2307/1170535

Barringer, C., & Gholson, B. (1979). Effects of type and combination of feedback upon conceptual learning by children: Implications for research in academic learning. *Review of Educational Research*, *49*(3), 459. doi:10.2307/1170140

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

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1850205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*, 633-642.

Bohlmann, N. L., & Fenson, L. (2005). The effects of feedback on perseverative errors in preschool aged children. *Journal of Cognition and Development, 6*, 119-131. doi:10.1207/s15327647jcd0601_7

Brainerd, C. J. (1972). Reinforcement and reversibility in quantity conservation acquisition. *Psychonomic Science, 27*, 114-116. doi:10.3758/BF03328907

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616. doi:10.3758/MC.36.3.604

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*, 290–298. doi:10.1037/a0031026

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology. Applied, 13*, 273–81. doi:10.1037/1076-898X.13.4.273

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281. doi:10.3102/00346543065003245

Chesney, D. L., McNeil, N. M., Matthews, P. G., Byrd, C. E., Petersen, L. A., Wheeler, M. C., Fyfe, E. R., & Dunwiddie, A. E. (2014). Organization matters: Mental organization of addition knowledge relates to understanding math equivalence in symbolic form. *Cognitive Development, 30,* 30-46. doi:10.1016/j.cogdev.2014.01.001

Christianson, K., Mestre, J. P., & Luke, S. G. (2012). Practice makes (nearly) perfect: Solving 'students-and-professors' type algebra word problems. *Applied Cognitive Psychology, 26*, 810-822. doi:10.1002/acp.2863

Corbett, A. T., & Anderson, J. R. (1989). Feedback timing and student control in the LISP intelligent tutoring system. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), *Proceedings of the Fourth International Conference on Artificial Intelligence and Education* (pp. 64-72). Amsterdam, Netherlands: IOS Press.

Coyle, T. R., & Bjorklund, D. F. (1997). Age differences in, and consequences of, multiple- and variable-strategy use on sort-recall task. *Developmental Psychology, 33*, 372-380. doi:10.1037/0012-1649.33.2.372

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Dihoff, R., Brosvic, G., Epstein, M., & Cook, M. (2003). The role of feedback during academic testing: The delay retention test revisited. *The Psychological Record, 53*, 533-548.

Evans, M. A., Barraball, L., & Eberlee, T. (1998). Parental response to miscues during child-to-parent book reading. *Journal of Applied Developmental Psychology, 19*, 67-84. doi:10.1016/S0193-3973(99)80028-8

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

Frantom, C. G., Green, K. E., & Hoffman, E. R. (2002). Measure development: The children's attitudes toward technology scale (CATS). *Journal of Educational Computing Research, 26*, 249-263. doi:10.2190/DWAF-8LEQ-74TN-BL37

Fyfe, E. R., & Rittle-Johnson B. (in press). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology.* doi:10.1037/edu0000053

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2015). When feedback is cognitively-demanding: The importance of working memory capacity. *Instructional Science, 43*, 73-91. doi:10.1007/s11251-014-9323-8

Fyfe, E. R., & McNeil, N. M, & Borjas, S. (2015). Benefits of "concreteness fading" for children's mathematics understanding. *Learning and Instruction, 35,* 104-120. doi:10.1016/j.learninstruc.2014.10.004

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104*, 1094-1108. doi:10.1037/a0028389

Gaddes, W. H., & Crockett, D. J. (1975). The Spreen-Benton Aphasia Tests: Normative data as a measure of normal language development. *Brain and Language, 2*, 257–280. doi:10.1016/S0093-934X(75)80070-8

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback on learning. *Learning and Instruction, 20*, 304-315. doi:10.1016/j.learninstruc.2009/08.007

Gilman, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assisted instruction. *Journal of Educational Psychology, 60*, 503–508. doi:10.1037/h0028501

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX: Results of experimental and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: North Holland.

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249-271). New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112. doi:10.3102/003465430298487

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help? *Journal of Experimental Child Psychology, 107*, 15–30. doi:10.1016/j.jecp.2010.03.004

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17*, 797–801. doi:10.3758/PBR.17.6.797

Hofer, S., Nussbaumer, D., & Schneider, M. (2011). Practice without feedback can increase the adaptivity of strategy choices. In J. Torbeyns (chair), *Strategy flexibility: Analyzing its related structures and processes.* Symposium presented at the EARLI Conference, UK.

Hoover-Dempsey, K. V., Battiato, A. C., Walker, J. T., Reed, R. P., DeJong, J. M., & Jones, K. P. (2001). Parental involvement in homework. *Educational Psychologist, 36*, 195-209. doi:10.1207/S15326985EP3603_5

Jacobs, P. I., & Kulkarni, S. (1966). A test of some assumptions underlying programmed instruction. *Psychological Reports, 18*, 103-110.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509-539. doi:10.1007/s10648-007-9054-3

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, 96, 558–568. doi:10.1037/0022-0663.96.3.558

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23–31. doi:10.1037/a0022243

Kamins, M. L., & Dweck, C. S. (1999). Person versus process praise and criticism: Implications for contingent self-worth and coping. *Developmental Psychology, 35*, 835-847. doi:10.1037/0012-1649.35.3.835

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*, 651-672. doi:10.1007/s11251-012-9209-6

Karabenick, S. A., & Knapp, J. R. (1988). Effects of computer privacy on help seeking. *Journal of Applied Social Psychology, 16*, 461-72.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis.* New York, NY: Guilford Press.

Kenny, D. A., Kashy, D. A., Mannetti, L., Pierro, A., & Livi, S.(2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83*, 126–137. doi:10.1037//0022-3514.83.1.126

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

Kieran, C. (1992). The early learning of algebra: A structural perspective. In S. Wagner & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 33-56). Hillsdale, NJ: Erlbaum.

Kluger, A. N., & Adler, S. (1993). Person- versus computer-mediated feedback. *Computers in Human Behavior, 9*, 1-16.

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284. doi:10.1037/0033-2909.119.2.254

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*, 67–72.

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.

Krause, U.-M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction, 19*, 158-70. doi:10.1016/j.learninstruc.2008.03.003

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211-232. doi:10.2307/1170128

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*, 505-512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279-308. doi:10.1007/BF01320096

Kulik, J. A., and Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*, 79-97.

Langer, E. J. (2000). Mindful learning. *Current Directions in Psychological Science, 9*, 220-223. doi:10.1111/1467-8721.00099

Luwel, K., Foustana, A., Papadatos, Y., & Verschaffel, L. (2011). The role of intelligence and feedback in children's strategy competence. *Journal of Experimental Child Psychology*, *108*, 61-76. doi:10.1016/j.jecp.2010.06.001

Mason, J. R., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What research tells us*. Retrieved from University of Nebraska-Lincoln, Center of Instructional Innovation from http://dwb4.unl.edu/dwb/Research/MB/MasonBruning.html

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

Matthews, P. G., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43*, 220-254.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist, 63*, 760–769. doi:10.1037/0003-066X.63.8.760

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43–52.

McGilly, K., & Siegler, R. S. (1989). How children choose among serial recall strategies. *Child Development, 60*, 172-182. doi:10.2307/1131083

McNeil, N. M. (2005). *When the basics backfire: The change-resistance account of equation-learning difficulties.* Unpublished dissertation, University of Wisconsin, Madison.

McNeil, N. M. (2008). Limitations to teaching children 2 + 2 = 4: Typical arithmetic problems can hinder learning of mathematical equivalence. *Child Development, 79*, 1524-1537. doi:10.1111/j.1467-8624.2008.01203.x

McNeil, N. M., & Alibali, M. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*, 883-899. doi:10.1111/j.1467-8624.2005.00884.x

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

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

McNeil, N. M., Chesney, D. L., Matthews, P. G., Fyfe, E. R., Petersen, L. A., & Dunwiddie, A. E. (2012). It pays to be organized: Organizing arithmetic practice around equivalent values facilitates understanding of math equivalence. *Journal of Educational Psychology, 104*(4), 1109-1121. doi:10.1037/a0028997

McNeil, N. M., Grandau, L., Knuth, E., Alibali, M. W., Stephens, A., Hattikudur, S., & Krill, D. (2006). Middle-school students' understanding of the equal sign: The books they read can't help. *Cognition and Instruction*, *24*, 367–385. doi:10.1207/s1532690xci2403_3

Miyake, A., & Shah, P. (1999). *Models of WM: Mechanisms of active maintenance and executive control*. New York: University Press.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, 32, 99–113. doi:10.1023/B:TRUC.0000021811.66966.1d

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd. pp. 745-783). Mahwah, NJ: Erlbaum.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner, & R. Brunken (Eds.), *Instructional design for multimedia learning* (pp. 181-195). Munster, NY: Waxman.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. *Learning and Instruction*, *16*, 310–322. doi:10.1016/j.learninstruc.2006.07.003

National Council of Teachers of Mathematics. (2014). *Principles to action: Ensuring mathematical success for all*. Reston, VA: NCTM.

Nihalani, P. K., Mayrath, M., & Robinson, D. H. (2011). When feedback harms and collaboration helps in computer simulation environments: An expertise reversal effect. *Journal of Educational Psychology*, *103*, 776–785. doi:10.1037/a0025276

Nussbaumer, D., Schneider, M., & Stern, E. (2014). The influence of feedback on the flexibility of strategy choices in algebra problem solving. *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1102-1107). Quebec City, Canada: Cognitive Science Society.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63-71. doi:10.1207/S15326985EP3801_8

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. doi:10.1037/0278- 7393.31.1.3

Petersen, L. A., & McNeil, N. M. (2013). Effects of perceptually rich manipulatives on preschoolers' counting performance: Established knowledge counts. *Child Development, 84*, 1020–1033. doi:10.1111/cdev.12028

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525–556. doi:10.3102/00346543074004525

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms [Supplemental material]. *Science, 315*, 1795-1796. doi:10.1126/science.1139719

Powell, S. R. (2012). Equations and the equal sign in elementary mathematics textbooks. *The Elementary School Journal, 112*, 627–648.

Rey, G. D., & Buchwald, F. (2011). The expertise reversal effect: Cognitive load and motivational explanations. *Journal of Experimental Psychology: Applied*, *17*, 33-48. doi:10.1037/a0022243

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

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

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

Roper, W. J. (1977). Feedback in computer assisted instruction. *Programmed Learning and Educational Technology, 14*, 43–49.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*, 450-461.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. doi:10.1037/1082-989X.7.2.147

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 352-359.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*, 759-775. doi:10.1037/a0025140

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189. doi:10.3102/0034654307313795

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31-76). Hillsdale, NJ: Erlbaum.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24*, 86-97.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 36*, 80–95. doi:10.1037/a0017407

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling – Second Edition*. London: Sage Publishers.

Snow, R. E. (1996). Aptitude development and education. *Psychology, Public Policy, and Law, 2*, 536-560. doi:10.1037/1076-8971.2.3-4.536

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

Sweller, J., Ayres, P., & Kalyuga, S. (2011). The redundancy effect. *Cognitive Load Theory*. (pp. 141-154). New York: Springer.

Sweller, J., Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Tobias, S. (2010). The expertise reversal effect and aptitude treatment interaction research (Commentary). *Instructional Science, 38*, 309–314. doi:10.1007/s11251-009-9103-z

Tudge, J., Winterhoff, P., & Hogan, D. (1996). The cognitive consequences of collaborative problem solving with and without feedback. *Child Development, 67*, 2892-2909. doi:10.1111/j.1467-8624.1996.tb01894.x

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

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – 4th Edition (WISC-IV)*. San Antonio, TX: Harcourt Assessment.

Wentling, T. L. (1973). Mastery versus nonmastery instruction with varying test item feedback treatments. *Journal of Educational Psychology, 65*, 50–58.