Toward Sense Making with Grounded Feedback

Eliane Stampfer Wiese

CMU-HCII-15-104
September 2015

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

Thesis Committee
Kenneth R. Koedinger (Chair)
Vincent Aleven
Aniket Kittur
Robert Siegler
Daniel L. Schwartz (Stanford University)

Submitted in partial fulfillment of the requirements
For the Degree of Doctor of Philosophy

Abstract

In STEM domains, robust learning includes not only fluency with procedures, but also recognition and application of the conceptual principles that underlie them. Grounded feedback is one instructional approach proposed to help students integrate conceptual knowledge into their learning of procedures. Grounded feedback functions primarily by having students take an action in the target domain (often symbolic) and receiving feedback in a representation that is easier to reason with. This thesis defines grounded feedback and evaluates its effectiveness.

I define grounded feedback with four characteristics: (1) The feedback reflects students' inputs according to rules that are inherent to the topic of study. For example, an inputted equation with two variables may be shown as a graph. (2) The feedback facilitates self-evaluation - by examining the feedback, students can evaluate for themselves if their answers are correct or not. (3) Students do not directly manipulate the feedback representation. Instead, the inputs are in a format that matches the domain learning goals. (4) The feedback conveys information about the nature of errors, not just that a particular action was incorrect. For example, the feedback may indicate the direction or magnitude of the error.

Some prior experiments on systems with the four characteristics of grounded feedback found greater learning of target procedures (Nathan 1998) and greater transfer (Mathan & Koedinger 20015), relative to robust controls. Over four studies with 4th and 5th graders, this thesis explores three tutor designs for fraction addition that incorporate visualizations of magnitude, including grounded feedback. Two studies of grounded feedback show effects of robust learning relative to correctness feedback, including greater future learning (in study 2) and transfer (in study 3). Another study found little difference between grounded feedback with and without correctness. In the last study, relative to correctness feedback, two implementations of dynamically linked concrete representations (variations on grounded feedback) showed greater robust learning (pre-test to delayed test). The correctness feedback tutor, used in three of these studies, is a high-bar control, including immediate step-level correctness feedback and adaptive on-demand hints. Indications of more robust learning with the grounded feedback tutor are promising, though not conclusive.

Grounded feedback is intended to leverage concrete representations to elicit students' prior knowledge of relevant concepts. Over two Difficulty Factor Assessments, 5th graders demonstrated difficulty incorporating magnitude information when evaluating fraction addition equations. In particular, students could generally evaluate an equation correctly when it was represented with fraction bars. However, including symbols with the bars interfered with students' evaluations by triggering incorrect transfer from whole-number addition. Students also did not fully grasp that when two positive fractions are added, the resulting sum is bigger than each addend alone. These findings may help explain why the benefits of grounded feedback are not as strong as proponents of concrete representations might hope. Namely, the target population may not be able to take full advantage of the magnitude visualization because they lack pre-requisite knowledge of how fraction addition involves magnitude.

Acknowledgements

Table of Contents

1 Introduction: Grounded Feedback

Summary. How can we design feedback that supports students in connecting steps in a procedure to conceptual principles? This chapter examines one approach that is present in many computer-based educational activities, but has been under-studied. Based on common characteristics, these systems are grouped together and the approach is termed "grounded feedback." In grounded feedback, students' inputs are in the target, to-be-learned representation, and their actions are reflected in a linked representation that is intrinsic to the domain and more familiar to students. To highlight the unique features of grounded feedback, I contrast it with similar instructional approaches for supporting sense making: correctness feedback, manipulatives, and linked representations. While experiments comparing grounded feedback to other approaches are limited, robust benefits of grounded feedback over correctness feedback are promising, and indicate that grounded feedback warrants further investigation.

1.1 Feedback for Sense Making

How can feedback best support learning in science, technology, engineering, and math (STEM) domains? In STEM domains, robust learning includes not only fluency with procedures, but also recognition of the conceptual principles that underlie them, and the appropriate application of those concepts (Schoenfeld, 1988). One obstacle to robust learning may be the notation in which these domains are communicated. Math, for example, is a language of its own with arbitrary conventions (e.g., '-' means 'subtract', but '=' does not mean 'subtract twice'). Students may learn the concepts

underlying these abstract symbols more easily when the unfamiliar symbols are connected to already-familiar representations that make relevant features more salient and, thus, are easier to reason with. For example, a student may think that 9/10 equals 11/12, since the second fraction is obtained by adding two to the numerator and denominator of the first fraction. When the two fractions are plotted on a number line, their magnitudes become more salient, and the student may reason that the two fractions are not equivalent.

Using multiple or non-symbolic representations is not a new idea. However, *how* to use these representations is still an open question. One promising approach is to use representations that support qualitative thinking, such as strip diagrams in Singapore Math (Fig. 1.1; Beckmann, 2004). Such diagrams are not intended to help students execute symbolic procedures, but rather are intended to support students in qualitative reasoning (e.g., Which amounts are bigger?) and planning (e.g., Which operation is needed?). *Grounded feedback* extends this idea of using an additional representation to support reasoning. Grounded feedback functions primarily by having students take an action in the target domain (often symbolic) and receiving feedback in a representation that is easier to reason with. The first representation is less familiar, and having students act in that representation forces students to engage with the target content. The second, easier-to-reason-with feedback representation is hypothesized to help students think conceptually about the problem. The link between the target representation and the feedback representation is hypothesized to help students see the connections between the less-familiar target representation and the underlying concepts of the domain. Grounded feedback is intended to support conceptual reasoning, even in drill-and-practice environments. To highlight the features of grounded feedback, I contrast it with similar approaches: correctness feedback, manipulatives, linked representations, and situational feedback. Experiments comparing grounded feedback to correctness feedback have shown benefits for grounded feedback; I have not found experiments comparing grounded feedback directly to the other approaches.

Orly spent 2/7 of her money on a pumpkin.
The pumpkin cost $10.
How much money did she have at first?

$10

?
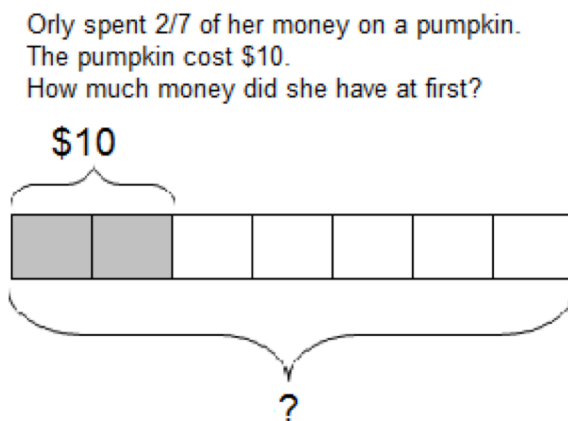
Fig. 1.1 Strip diagrams support qualitative reasoning (Orly had more than $10 at first) and planning (find half of $10, then multiply that amount by 7 to find the original amount). However, the diagram does not directly support using abstract symbolization to solve the problem.

1.1.1 Defining Grounded Feedback

Grounded feedback as I define it has four criteria, which are domain-general:

1) The feedback is intrinsic to the domain and semantically equivalent to the students' inputs. The connection between the input and the feedback is governed by rules that are inherent to the topic of study. For example, an inputted equation with two variables may be shown as a graph.

2) Students can easily envision the feedback state that indicates a correct answer to a given problem. Therefore, by examining the feedback, students can evaluate for themselves if their answers are correct or not. Further, the feedback representation must facilitate evaluation: the task of deciding if a problem has been solved correctly or not must be easier with the inclusion of the feedback representation than without it (that is, just looking at the inputs alone). This feature depends on the prior knowledge of the students.

3) Students do not directly manipulate the feedback representation. Instead, the inputs are in a format that matches the domain learning goals. This design feature ensures that students engage with the new representation.

4) The feedback affords inferences on errors. The feedback conveys information about the nature of errors, not just that a particular action was incorrect. For example, the feedback may indicate the direction or magnitude of the error.

1.1.2 Darts: An Example of Grounded Feedback

Dugdale's Darts program (1992) is one example of grounded feedback (Simcalc is another – see Roschelle, Kaput, & Stroup, 2000). Students attempt to pop balloons on a number line by shooting darts, aimed by entering a number or expression. The dart flies to that location on the number line and stays there, with the original numeric input beside it (Fig. 1.2). If a dart touches a balloon, the balloon pops. This example illustrates the four characteristics of grounded feedback:

1) *The feedback is intrinsic to the domain.* The placement of the dart on the number line is governed by the underlying mathematics, and the same magnitude information is conveyed in both the symbolic and graphical representations. The feedback is *intrinsic* because there is a consistent mapping from the inputs in the target representation to the displayed responses in the feedback representation (Dugdale, 1992). Intrinsic feedback shows students the workings of the domain, and therefore is the same whether or not the input is correct: A dart will always land at its specified location on the number line, even if no balloon is there.

2) *Students can envision a correct goal state for the feedback.* A correct input will result in a dart that lands in the same location as the balloon the student is aiming for, and will pop the balloon. This indication of correctness is quite explicit, yet the feedback is more than simple verification.

3) *The input format matches the domain learning goals.* Students input numbers and expressions instead of, say, popping the balloons by clicking on them.

4) *The feedback affords inferences on errors.* This feedback conveys more information than simply if the input has the same magnitude as the target balloon. By comparing the dart's location to that of the target balloon, the student can tell if a larger or smaller number is needed. Further, the feedback representation facilitates a rich set of inferences, for example, the student could infer that "1/3 + 1/6" is about halfway between 1/3 and the balloon target above, and thus infer that "1/3 + 2/6" might be a good next entry.
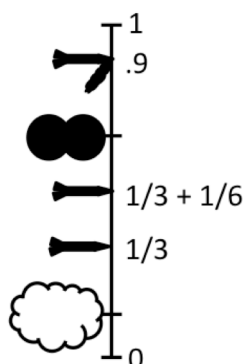
Fig. 1.2 Sample reconstructed screenshot illustrating the feedback in Darts with a number line from 0 to 1. The student has already popped a balloon at .9.

1.1.3 Theoretical Context

Grounded feedback is grounded both in a different representation and in student's prior knowledge. Darts dynamically translates a student's intention (estimate a balloon's location with a number or expression) to a representation (the number line) that may help the student see if that action matched the original intention (is the target balloon located at that spot?). It lets the system ask "is this what you mean?" perhaps giving pause when the feedback is not what the student expected. Grounded feedback environments allow students to iterate on their work through cycles of generation and feedback. Through this iteration, the feedback encourages students to evaluate their own work, and the act of evaluation may deepen conceptual knowledge.

Cognitive tutors often give immediate *explicit feedback* that tells students whether an answer is correct, for example by changing the color of incorrect answers to red. The tutor considers any input that is not on a solution path to be incorrect. Ohlsson, (1996) termed this interpretation of errors the "objective view of errors." Novices may not recognize objective errors on their own. However, novices can recognize when the consequences of an action violate their expectations, which

Ohlsson called the "subjective view of errors" (Ohlsson, 1996). Students can learn from subjective errors when they adjust their mental model to reconcile their previous understanding with the actual results of their actions. Grounded feedback provides the context in which these subjective errors can occur. If students can recognize subjective errors, why do they make them in the first place? Ohlsson (1996) hypothesizes that the skills needed to produce versus evaluate correct actions are distinct and may draw on different knowledge bases (that is why we often correct mistakes when we check our own work, even without new information). Therefore, when students evaluate their own work, they may be activating knowledge that was not available while they produced the answer. Such evaluation may help with knowledge acquisition by reminding students of what they already know and prompting them to engage with what they think is true. Grounded feedback may help by making that evaluation step more explicit and by providing information, which a novice can interpret, about why their action was in error (Powers, 1973). Other supports for self-evaluation, such as prompted self-explanations, are also intended to support sense making, but grounded feedback is distinguished in being a reflection of student actions rather than an instructional prompt.

Feedback that flags actions as correct or incorrect also facilitates evaluation – by doing it for the student. When such feedback indicates an action is incorrect, the student learns a negative example. Grounded feedback not only shows errors as negative examples for that step, but also as positive examples for something else. For example, the dart at 1/3 + 1/6 provides a negative example for the balloon (the balloon is not at 1/3 + 1/6) and provides a positive example for 1/3 + 1/6 (showing where it is located on the number line). Since the goal is for students to understand how the novel representation maps to the more familiar one, each example of mapping is an opportunity for learning, even if the input being mapped is incorrect for that particular problem or step. In addition to using incorrect inputs as positive examples, grounded feedback allows for a rich set of inferences based on those mistakes: 1/3 + 1/6 is smaller than the two-part balloon; it is about halfway between 1/3 and the two-part balloon; 2/3 + 1/6 might hit the two-part balloon, etc. Although some of these inferences could be given as text feedback, enumerating all of them would be overwhelming. Further, with grounded feedback the students actively make these inferences themselves.

Grounded feedback uses two representations for a specific aim: to show feedback in a familiar representation that facilitates students' evaluation of their own work in the novel target representation. Multiple representations in general can have much broader aims and different types of structures. For example, the goal for an activity with multiple representations may be for students to discover the mapping from one to another, to construct one from the other, or to use each separately. This diversity in types and uses of multiple representations helps explain why research in this area can appear to show contradictory results: Research on one use of multiple representations may not generalize to other uses (Ainsworth, 1999). Ainsworth (1999) described three categories of functions of multiple representations: (1) providing complimentary information (e.g., different map projections of the earth, one with accurate shapes and the other with accurate sizes); (2) constraining potential misinterpretations (e.g., a

linked representation of an object moving according to a velocity-time graph to address a common misconception that a horizontal line indicates an object at rest); and (3) constructing deeper understanding (e.g., combining base-10 blocks with symbolic numbers to encourage students to extract principles of the base-10 number system). Grounded feedback has elements of the second and third categories: The grounded representation facilitates correct interpretation of the novel representation, and by mapping between both representations, students construct deeper understanding of the domain. One goal in defining grounded feedback is to indicate a particular kind of multiple representation use that may be effective in supporting learning of procedures with understanding.

Grounded Feedback is also a type of formative feedback (Shute, 2008). Shute suggests several dimensions for categorizing feedback. Facilitative feedback helps "guide students in their own revision and conceptualization" (Shute, 2008, p. 157) while directive feedback "tells the student what needs to be fixed or revised" (Shute, 2008, p. 157). Two other related categories are verification (right/wrong feedback), and elaboration, which is "the informational aspect of the message, providing relevant cues to guide the learner toward a correct answer" (Shute, 2008, p. 158). Grounded feedback gives implicit verification (the student can compare the grounded feedback to its expected goal state), elaboration (it shows the nature of errors), and is facilitative, helping students come to their own understanding of the domain through repeatedly seeing their input matched to a more familiar representation. Facilitative and elaborative feedback often leads to better learning than simple verification (Shute 2008). However, Shute's review focused on explicit feedback. Grounded feedback's elaborative aspects are implicit in students' interpretation of the grounded representation. A key goal of this chapter is to explore when students' interpretation of such representations may lead to better learning than their interactions with explicit feedback.

1.2 Contrasting Grounded Feedback with Similar Instructional Approaches

How is grounded feedback different from other common types of feedback? This section compares and contrasts grounded feedback with correctness feedback, manipulatives, and multiple representations (both linked and non-linked). Table 1.1 indicates which of the four features of grounded feedback are present in the other feedback types. The first two features are characteristics of the feedback design, and the last two features depend on the students' responses.

Some entries may appear counter-intuitive. For the non-linked representation, students' envisioning of the correct goal state is *Not Applicable*: Since the non-input representation is not linked to anything, it only displays one state and therefore there is nothing to envision. Additionally, the second representation does not directly give feedback on the input representation, so *feedback is intrinsic* is listed as *no*. However, students can sometimes use the non-linked representation to diagnose errors, so the last feature is listed as *sometimes*. The goal of this set of contrasts is to highlight the

features of grounded feedback. In several cases we hypothesize that grounded feedback may be more beneficial than the contrasted approach. However, in all of these comparisons, there is little or no empirical evidence for the superiority of one approach or the other. One goal of this review is to call for experiments that would provide such evidence.

	Design Features		Likely Student Responses	
	Input matches domain learning goals	Feedback reflects the inputs, and affords inferences on errors	Students can use the feedback to decide if their work is correct	Students can interpret the feedback on their errors
Grounded Feedback	Yes	Yes	Yes	Yes
Step-level Correctness	Yes	No	Yes	Yes
Manipulatives	No	Yes	Yes	Yes
Non-linked Representations	Sometimes	No	Not Applicable	Sometimes
Linked Representations, Other Direction	No	Somewhat	Unlikely	Limited
Situational Feedback	Yes	Yes	Sometimes	Sometimes

Table 1.1 Comparing grounded feedback to other instructional approaches.

1.2.1 Verbal and Correctness Feedback in Tutoring Systems

How is grounded feedback different from correctness feedback? Consider the correctness feedback provided by the Algebra Cognitive Tutor (Koedinger & Aleven, 2007) shown in Fig. 1.3. The tutor marks incorrect inputs, such as *.13t* in the second column, and prevents students from progressing in the problem until that error is fixed. For recognized common errors, the system provides text feedback. Here, the tutor explains that the student's expression for the cost of t minutes of phone calls, *.13t*, does not include the base charge of $14.95 per month. While several features of this learning environment match the requirements of grounded feedback, one does not. Starting with features in common: students' inputs are in a representation that

matches the domain learning goals; students can envision a correct goal state for the feedback (if inputs change to green, they are correct); and students can interpret the feedback the system provides (green means correct, red means wrong).

My current cell phone company charges me $14.95 per month for service and $.13 per minute. PPS Cellular Phone Company has offered me $15.00 worth of free calls a month if I switch, but the charge is $.39 per minute.

1. How many minutes of calls can I get from PPS Cellular Phone Company for $50? What is the cost from my current company for that number of minutes?

Quantity Name	Time	Current cost	PPS cost
Unit	minutes	$	$
Expression	t	.13t	
Question 1			

The cost from my current company increases by 0.13 each minute, but remember that it starts at 14.95 dollars.

Fig. 1.3. Reconstructed sample work and feedback with the algebra cognitive tutor. The student's symbolization *.13t* is immediately marked as wrong, without giving the student the opportunity to evaluate his own work.

The key difference between grounded feedback and the Algebra Cognitive Tutor feedback is that correctness feedback is not intrinsic to the domain and thus does not promote the same degree of inferences on the nature of the students' errors. Fig. 1.4 shows one possible redesign to make this tutor grounded. First, before constructing the expression, students calculate the cost of the current cell phone plan for various numbers of minutes. While it may seem counter-intuitive, it is actually easier for novices to provide a numerical answer to a story problem than it is to construct the symbolic expression that yields that answer (Heffernan & Koedinger, 1998). To guard against slips, students could get correctness feedback on the calculated costs. After finding the correct costs, students propose an expression, which the tutor evaluates for each of the given number of minutes. This allows the students to judge the correctness of the expression by comparing its results to the correct cost that they just calculated. In the grounded feedback version, students can see what the costs would be if the charges were only 13 cents per minute. By comparing the values derived from the expression to the ones they calculated, the students are likely to determine for themselves if they have made an error. Upon considering the zero-minute row, this student is likely to see for himself which part of the expression he forgot.

Quantity Name	Time	Current cost (calculated)	Current cost (expression)
Unit	minutes	$	$
	0	14.95	0
	1	15.08	.13
	2	15.21	.26
	10	16.25	1.30

Expression for your current cost:

.13t

Fig. 1.4. Proposed implementation for grounded feedback in an algebra tutor, showing the same student error as Fig. 1.3. First, the student calculates the charges for various numbers of minutes. After inputting an expression for the current cost, the computer uses that expression to generate the costs for the given numbers of minutes. This allows the student to evaluate her own work, and to see what costs are generated by incorrect expressions.

Intelligent tutoring systems use immediate correctness feedback to reduce the unproductive cognitive load that comes from floundering. However, as Fig. 1.4 shows, grounded feedback can be implemented in a way that is compatible with cognitive load theory. Since grounded feedback is calibrated to students' prior knowledge, the tasks and decisions students are given are ones for which the students have a high probability of success. This is in line with cognitive load theory design principles, which propose that students should perform problem-solving steps (including evaluating their work) if they are likely to do so efficiently. Experiments comparing grounded to correctness feedback have found benefits for grounded feedback (see section 1.3, 'Evidence on the Effectiveness of Grounded Feedback').

1.2.2 Manipulatives

Like grounded feedback, manipulatives are intended to support student inference and sense making. For example, when solving a problem such as *20 − 7* with blocks, a student would see that they cannot take away seven unit blocks from two ten blocks (Fig. 1.5a). Trading one of the ten blocks for ten unit blocks (Fig. 1.5b) makes the subtraction possible: take seven of those unit blocks away, and 13 units remain (Fig. 1.5c). This use of manipulatives is intended to show the conceptual basis for borrowing.

While student use of manipulatives involves cognitive processes that are *analogous* to the target cognitive processes (e.g., borrowing with blocks is analogous to borrowing in the place value representation of multi-digit numbers), such use does not directly require and may not involve those target processes (Sarama & Clements, 2009). In other cases, the strategy that students practice may not generalize to all

problem types (e.g., picture-division strategies for fraction division are ill-suited to numbers that do not divide evenly; Rittle-Johnson & Koedinger, 2001). Further, in some cases students may not even realize what the manipulatives are intended to model - for example, students moving a token a set distance along a number line may not realize their actions are intended to model addition (Suh, Moyer, & Heo, 2005).
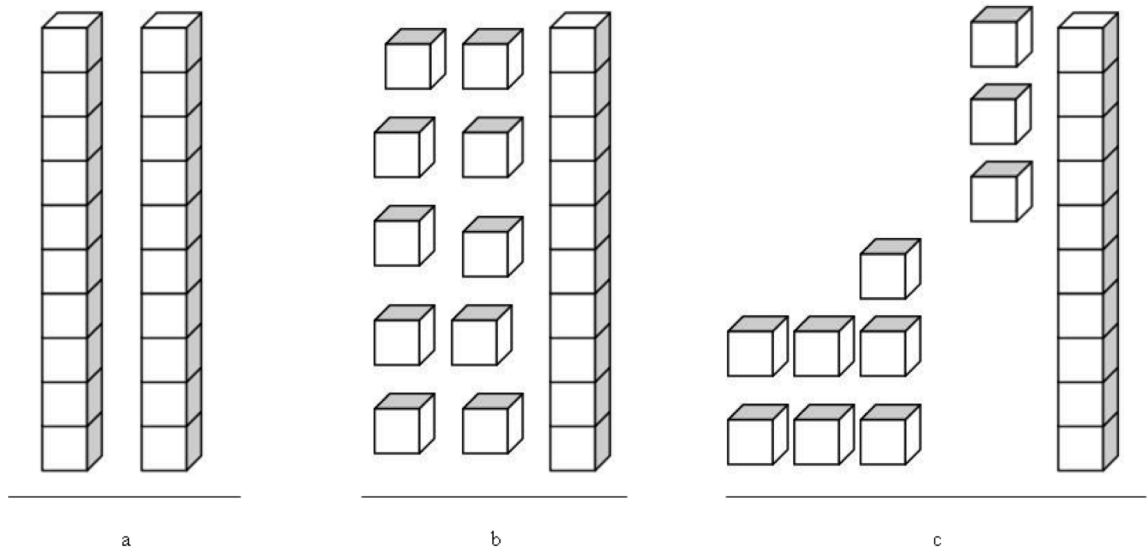
Fig. 1.5. Solving 20 minus 7 with base-ten blocks: a) 2 sticks of 10; b) trading one stick of 10 for 10 unit blocks; c) taking away 7 of the unit blocks.

Additional evidence that manipulatives do not always involve target cognitive processes is the lack of transfer between learning with manipulatives and posttests with paper and pencil. In a study on subtraction with base ten blocks, Resnick & Omanson (1987) found that students' ability to solve subtraction problems with the blocks was not predictive of proficiency with paper-and-pencil subtraction problems. While students seemed to learn subtraction concepts with the blocks, the procedures they acquired did not directly transfer to the multi-digit number representation. It seems the concepts they may have acquired were either not flexible enough or required too much cognitive load for novices to adapt them to numbers. From a situated cognition perspective (Lave, 1988), manipulatives produce knowledge that is tied to the use and affordances of the manipulatives and will thus be difficult to access without them. Uttal et al. (2013) found that this problem of transfer between manipulatives and paper-and-pencil tasks works both ways: rising second graders taught two-digit subtraction with one method performed significantly worse when tested in the other method (adjusted mean scores for posttests in the other method were less than 40%, while adjusted mean scores for the same method were above 85%). Together, this set of work indicates that manipulatives do not provide full support for students to connect concepts to procedures with abstract representations.

In contrast to manipulatives, with grounded feedback students are required to engage with the target representation, with the aim that the procedures and concepts they acquire are situated within that representation. I hypothesize that students will

primarily use the input representation to generate responses and the feedback representation to check their work. Therefore, as students develop competence with the new representation, errors should reduce and self-evaluation should get faster. Thus, the need to attend to the grounded feedback representation should fade. This progression contrasts with manipulatives, which need to be actively withdrawn. While I hypothesize that grounded feedback is more beneficial than manipulatives, I am not aware of any studies that compare the two.

1.2.3   Non-Linked Representations

This section discusses two systems that use non-linked representations, as they are similar to grounded feedback. One uses physical models as the more-familiar representation (similar to Izsak, 2000), and the other is entirely on the computer. In both cases, students work primarily in the less-familiar representation and use the more-familiar representation to check their work. Unlike grounded feedback, the two representations are not linked. Instead of showing the student's current work, the familiar representation only shows the correct answer.

In Padalkar & Hegarty's chemistry instruction (2012), students started with a diagram of a molecule, and were asked to draw the same molecule with a different type of diagram. Students were also given a three-dimensional ball-and-stick model of the molecule. Students completed six such problems as a pretest. After the pretest, the intervention group was instructed in how to use the 3-D models to check their work (the other group got a five-minute break). During the instruction, students were told to map between the 3-D model and the given diagram in the problem to verify that they represented the same molecule. Then, they were told to map between the 3-D model and their generated diagram. If the mapping was not possible, that indicated that the student's solution was incorrect. After this brief instruction, students completed a matched post-test with molecules that were mirror images of the pretest molecules, and a transfer task with new molecules.

While the chemistry instruction required students to perform the mapping between the two representations, this mapping step was done automatically in the QUADRATIC tutor (Wood & Wood, 1999). In this tutor, students expand quadratic expressions such as $(x + 1)^2$. As students form symbolic expressions, the tutor maps correct components onto a geometric model (Fig. 1.6) Students can verify that their work is correct when all terms have been mapped. The main difference between these non-linked approaches and grounded feedback is that the more-familiar representation does not reflect the student's actions if those actions are incorrect. It is not clear which approach is better for learning, and I am unaware of any experiments that compare grounded feedback to non-linked representations. I offer two competing perspectives: (1) The non-linked approach is better because it encourages students to actively integrate the two representations; (2) The grounded approach is better because it helps students diagnose their own errors in mapping between the two representations.
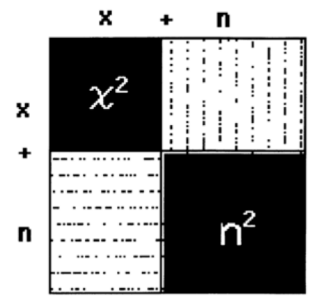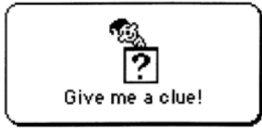
Fig. 1.6. In the QUADRATIC tutor, students expand expressions such as $(x+n)^2$. Correct terms are mapped to a geometric model.

To learn effectively from multiple representations, students must integrate them and know how to map between them (Ainsworth, Bibby, & Wood, 2002). In grounded feedback systems, a computer often does this mapping by generating a reflection of the student's actions in the feedback representation. In non-linked representations where a mapping is not provided, the student must do this mapping to evaluate her work. It is possible that this active student integration is a key ingredient for student learning with these systems, a view supported by evidence showing that students learn more when they integrate static representations rather than view pre-integrated ones (Bodemer, Plötzner, Bruchmüller, & Häcker, 2005). If active student integration improves learning, students would learn more from a non-linked system where the mapping is not provided than from either a non-linked system where mapping is provided or from grounded feedback.

On the other hand, students may not spontaneously map between the two representations, and may not do so correctly. Padalkar & Hegarty (2012) found that students who were explicitly taught how to use the 3-D models to check their work ended up using those models more on the post-test than they did on the pretest (the difference in the control group's usage of the models between the two tests was not significant). Further, the instruction group improved more from pre- to post-test than the control group (Padalkar & Hegarty, 2012). This indicates that on their own, students were not benefitting from the opportunity to actively integrate the two representations. Indeed, students in a pilot study often did not bother to check their

work at all because they were (over) confident in their answers (Padalkar & Hegarty, 2012). Even when students do map between the two representations on their own, they may do so incorrectly. For example, an algebra student trying to convey "40 subtracted from 800" may write "40 - 800," not realizing that the minus sign can only mean "subtract" and never "subtracted from" (Koedinger, personal communication). In that case, a familiar representation showing 760 would not help the student realize that her work was incorrect – she thinks that 40 - 800 does equal 760. Further, the unfamiliar representation itself may trigger misconceptions. For example, Roschelle, Kaput, & Stroup (2000) discuss students' misconceptions around an elevator simulation, where an elevator goes up and down at various speeds, represented by a piecewise linear function of velocity. While interpreting these graphs, students often confuse "going down" with "slowing down," thinking that a decrease in velocity represents the elevator moving downward, when it actually represents the elevator moving upward, but more slowly (Roschelle, Kaput, & Stroup 2000, p. 18). If the student is left to interpret these unfamiliar representations on her own, she may do so incorrectly and not realize it. If students are likely to map incorrectly or not at all, grounded feedback systems would be more beneficial than non-linked systems, as the student would see the familiar representation behave differently than expected, giving the student feedback on the meaning of the unfamiliar representation.

1.2.4   Linked Representations, from Concrete to Abstract

The Equivalent Fractions applet (illuminations.nctm.org/Activity.aspx?id=3510; Fig. 1.7) is one example of a linked representation that connects a familiar representation (fraction rectangles) to an unfamiliar representation (fraction symbols). The key difference between this feedback and grounded feedback is the direction of the link: grounded feedback uses the unfamiliar representation as input and the familiar one as feedback, while this applet does the reverse. This example illustrates the differences between the two link directions. The equivalence applet presents one fraction and asks the student to generate two equivalent fractions with different denominators. Each proper fraction n/d is represented in three ways: as a fraction rectangle with d equal pieces, n of which are colored in; as a location on a zero-to-one number line, and as a symbolic fraction. Students generate equivalent fractions by moving the horizontal and vertical sliders next to each fraction rectangle to create equal divisions. Next, the student clicks on the divisions to color them in. As the student manipulates the rectangles, the corresponding symbolic fractions show the numerator and denominator, and move along the number line as the student adjusts the number of colored-in pieces. When all of the fractions are equivalent, all three points will overlap on the number line. To confirm that the fractions are equivalent, the student can press the check button.
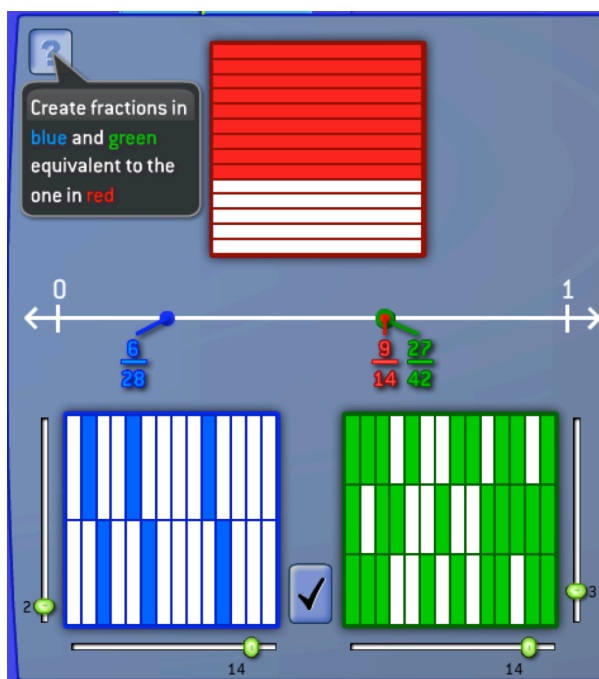
Fig. 1.7. Students input fractions by using the sliders to generate equal-sized pieces and clicking on each piece to color it in. The symbolic and number line representations are dynamically linked to the rectangles. The feedback is not grounded since students manipulate the more-familiar representation and get feedback in the unfamiliar representation.

What are the differences in design features between this type of linked representation and grounded feedback, and the likely differences in student learning? With the applet, students directly manipulate the more-familiar rectangle representation instead of the less-familiar symbolic fractions. All three representations reflect the underlying mathematics of fractions, and therefore the feedback is intrinsic to the domain. However, while the feedback representations afford some inferences on errors, they do not support as many types of inferences as the rectangle representation, which is the original input representation. The number line representation indicates if the fractions are equivalent or not and the magnitude differences between the fractions. However, these inferences would be available with the rectangle representation if all three rectangles were aligned and pieces were colored in consecutively.

Further, with the number line representation, students do not get feedback on whether they have an incorrect denominator, numerator, or both. If the rectangles were all aligned, they could show this type of information. Since the feedback representation is unfamiliar, it is unlikely that students will be able to envision the correct goal state for the feedback. In this example, students would probably be able to envision the correct goal state for the number line representation: the points should overlap. It is less likely that a novice would be able to envision the correct goal state for the fraction symbols. If they could, this lesson would most likely not be

necessary! Since students are unlikely to be able to determine if their work is correct from the fraction symbols alone, it is very unlikely that students would be able to use the symbols to generate meaningful inferences about their errors.

As with manipulatives, when students directly manipulate a more-familiar representation, they may be tempted to ignore the unfamiliar representation. With this applet, a student could generate equivalent fractions without considering the symbolic representations at all: the student could count the divisions in the red fraction, set the horizontal sliders to that number, set the vertical sliders to two different numbers greater than one, and color in pieces until the three points overlap on the number line. Though the applet may help students solidify equivalence concepts, it does not give students practice with procedures for generating equivalent fractions with symbols, and thus is unlikely to lead to robust learning of the procedures or robust understanding of how the procedures and concepts are connected.

One way to make this activity grounded is presented in Fig. 1.8. Students input symbolic numbers in the areas marked with black boarders. The equations encourage students to multiply, but the interface does not require it. If students choose to multiply the denominator, the original horizontal divisions are overlaid with vertical divisions. If the student chooses not to multiply, the rectangle shows only horizontal divisions. The number of colored pieces is driven by the student-entered numerator, and pieces are colored consecutively. Since the rectangles are aligned, students can compare the magnitudes without the number line. Alignment also should facilitate comparison of denominators (e.g., eighths do not line up with fourteenths). Although I hypothesize that grounded feedback will be more beneficial than familiar-to-unfamiliar linked representations, I hasten to add that I have not found experiments comparing these two designs. Linked representations from the more-familiar to the less-familiar is a popular design choice (e.g., phet.colorado.edu, shodor.org, nlvm.usu.edu), and evidence is needed on whether this is the most beneficial choice for learning.
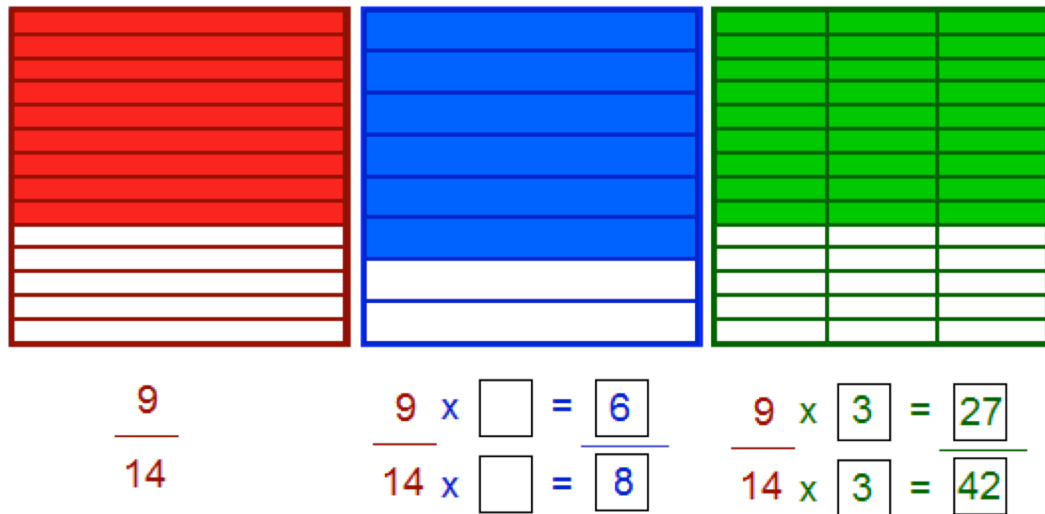
Fig. 1.8. Proposed re-design of equivalence applet that uses grounded feedback. Students enter symbolic numbers in the black-bordered input areas, and see the rectangle representation of the fractions as feedback. The equations encourage students to multiply, and to think about multiplication when they interpret the feedback. Since the fraction rectangles are aligned, students can compare their fractions without the number line.

1.2.5 Situational Feedback

Situational feedback draws on theories of a *problem model* and a *situation model* when students encounter story problems. While the *problem model* "represents the mathematical structures needed to solve the problem" (Nathan, 1998, p. 139), the *situation model* "draws on the reader's prior knowledge of events and semantic knowledge" (Nathan, 1998, p. 139). Experts draw on both models. When novices rely on the problem model to the exclusion of the situation model, they may generate nonsensical answers. For example, an 8th grade question from the third National Assessment of Educational Progress asked, "An army bus holds 36 soldiers. If 1128 soldiers are being bussed to their training site, how many busses are needed?" 29% of students answered "31 remainder 12" (Schoenfeld, 1988. p. 6). Situational feedback aims to help student connect the mathematics in the problem model to the situation model. Nathan's ANIMATE system (1998)is one implementation of situational feedback.

The ANIMATE tutoring system teaches students how to model a story problem with algebra equations. Students set up equations, which drive animations, which the student can then compare to the situation in the story. A sample problem: a train leaves its station going 75 miles per hour. A helicopter leaves from the same station two hours later, going 300 miles per hour, to warn the train that there is a broken

bridge 60 miles ahead. Can the helicopter catch up with the train in time? Fig. 1.9 shows a sequence of example student work and feedback for this problem.
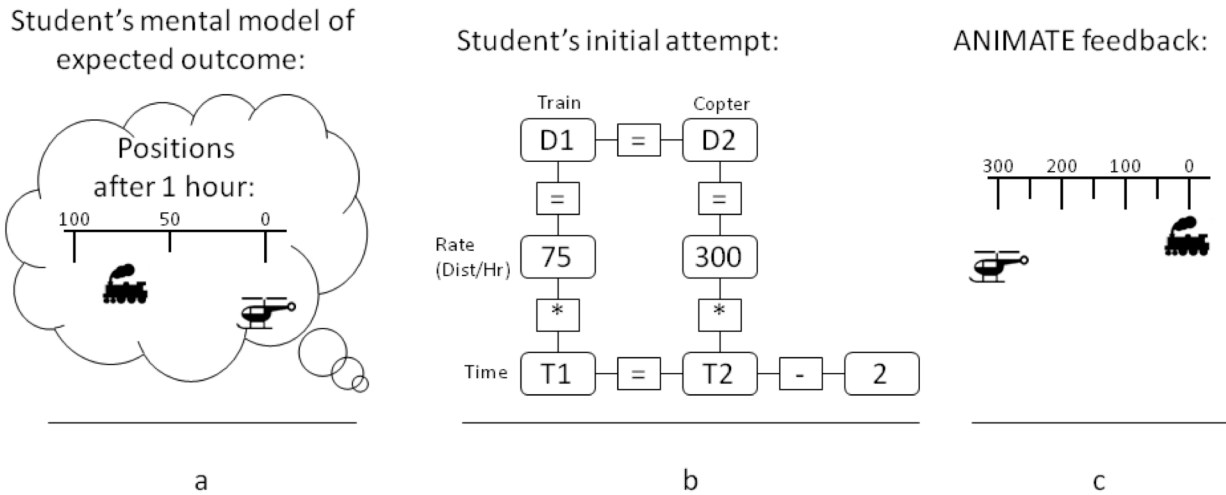
Fig. 1.9. a) Student's expectation that after one hour, the train will have gone 75 miles and the helicopter will not have left. b) Student's inputted equations for modeling the story problem. c) ANIMATE's feedback, based on the entered equations, does not match the student's expectations. The feedback supports qualitative reasoning, while the input format supports algebraic symbolization.

Fig. 1.9a shows the student's expectation that after an hour, the train will have traveled 75 miles and the helicopter will not have left the city yet. Fig. 1.9b shows the system of equations the student has entered to model the story problem: D1 = D2 (both vehicles have traveled the same distance once the helicopter catches up with the train); D1 = 75 * T1 (the train's distance is its speed multiplied by its travel time); D2 = 300 * T2 (likewise for the copter). T1 = T2 − 2 relates the amount of time that the two vehicles have been traveling, demonstrating a common misconception. The student tried to model that the copter leaves two hours after the train, perhaps thinking if the train left at 9am, the helicopter would have left at 11am, and 9 = 11 − 2. However, the equation requires that T1 and T2 represent the amount of time each vehicle has been traveling, not the clock time when they left. Therefore, the correct equation is "T1 = T2 + 2" since the train travels for 2 hours more than the helicopter. These entered equations drive animations of the train and the helicopter. Fig. 1.9c shows the animation for the positions of the train and helicopter after an hour. Unlike the student's expected outcome in Fig. 1.9a, Fig. 1.9c shows that the helicopter travelled 300 miles and the train stayed at the station. A full reconstruction of the ANIMATE interface at this point is shown in Fig. 1.10. In this example, the animation would show the chase helicopter leaving *before* the train, which does not match the problem. Ideally, the student reconsiders the equations he entered to locate the error.

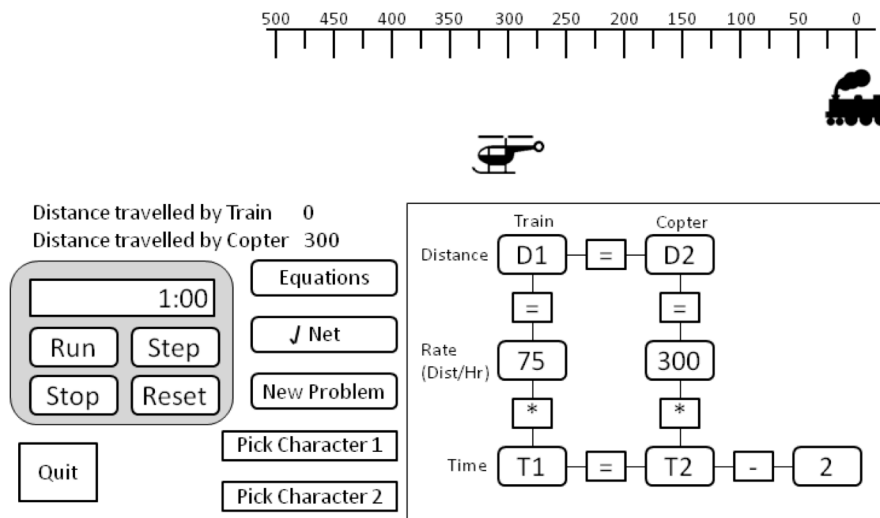**Fig. 1.10.** Reconstructed ANIMATE screenshot. The student's entered equations are in the box at the bottom right; the clock at the left shows one hour has been simulated; the animation at the top shows the movement of the train and helicopter as governed by the student's equations. The animation does not match the story description.

The ANIMATE system is both situational and grounded. The ANIMATE system has no student model, does not mark answers as correct or incorrect, does not provide text hints, and does not force students to attain the correct answer before moving on to the next problem. Instead, its key feature is providing student-meaningful situational feedback "in such a way that the learner can use her prior knowledge to identify solution errors, re-examine prior conceptions, and propose and test hypotheses about the causes of errors." (Nathan, 1998, p. 138). This is also a key feature of grounded feedback. However, it is not always present in situational feedback environments, for example in Horwitz and Barowy's RelLab (1994).

The RelLab simulation environment (Horwitz & Barowy, 1994) is an example of situational feedback that is not grounded. RelLab was designed to teach physics concepts to high school students, and performs simulations of events at normal and relativistic speeds. In many ways RelLab is similar to ANIMATE: students read a scenario, input parameters to a computer program, and watch an animation play out. However, since RelLab's problem scenarios often challenge students' preconceptions, the animations are not sufficient feedback for students to know that they set up the problem incorrectly, violating the second criterion of grounded feedback. When the target content involves conceptual change, students are unlikely to correctly envision the goal state, meaning that they cannot effectively use their expectations to evaluate the situational feedback. Indeed, while RelLab was successful in prompting discussions and helping students understand physics, Horwitz & Barowy describe cases where the RelLab animations were not sufficient to alert students to errors,

because the students did not have a prior conception of what a correct animation would look like (1994).

RelLab illustrates how situational feedback is different from grounded feedback. Situational feedback is a property of the learning environment, but grounded feedback is a property of the match between the learning environment and students' prior knowledge. Our point is not to say that RelLab is a poor learning environment, but rather that story problems and feedback given in a situated context does not necessarily make it easier for students to tell if their work is correct. Conversely, Darts is an example of a system that is grounded but not situational. While Darts does involve grounded feedback, it does not involve a story problem or require the student to generate a "situation model" based on semantic relationships and prior knowledge of events. Therefore, I argue that grounded feedback and situational feedback may overlap, but one is not necessary or sufficient for the other.

1.3   Evidence on the Effectiveness of Grounded Feedback

1.3.1   Animate: Comparing Grounded Feedback to Error Messages

ANIMATE (discussed in detail above) is one of the few grounded feedback systems that has been compared to a control, with learning measured with pre- and post-tests outside the tutors (Nathan, 1998). Instead of running simulations, the control tutor gave three pop-up hints when the student made an error (e.g., at the error depicted in Fig. 1.10, the first hint reads "It is common to over-generalize 'later than' to mean minus. Please check your current work.") An experiment with 31 college students using a pretest-intervention-posttest design showed that while both groups improved in modeling story problems from pre- to posttest, the situational feedback group improved more. The tests included one problem of each type: travel (example show in Fig. 1.10); investment (e.g., \$750 is invested at an interest rate of 5%, compounded annually. How much is in the account at the end of the second year?); and work (e.g., Tom can paint the entire fence in two hours while it takes Huck four hours. If Tom arrives one hour late from fishing, how long will it take the two boys to complete the job?). Separate ANCOVAs for each problem type with pretest performance and total SAT (the standardized test) as covariates and treatment as a between-subjects factor showed a significant difference for treatment ($p < .05$) on travel and investment problems, with .76 for the standardized gain for the situational feedback condition on both types and .57 and .43 for the control, respectively. The ANCOVA for work problems did not show a significant difference between conditions.

While Nathan's 1998 study shows strong overall benefits for situational/grounded feedback (relative to the pop-up text hints), student learning was not significantly different on work problems. Work problems differed from the other types in that students were given a whole number in the problem statement (Tom can paint 2 fences per hour) but they needed to use the reciprocal in the

equation (Tom needs 1/2 an hour per fence). ANIMATE students could see that using the original number was incorrect, but did not know what to try next. This finding suggests that grounded feedback may only be more beneficial than right/wrong feedback or pop-up text hints when students are able to both recognize that they have made an error and make useful inferences on those errors to guide them on what to try next. Alternatively, a system that provides both grounded feedback and text hints (when students are stuck) may be more powerful than either type of support on its own. Mathan and Koedinger's Excel tutor (2005) is one such system.

1.3.2 Excel Tutor: Comparing Grounded Feedback to Correctness Feedback

Mathan and Koedinger's 2005 Excel tutor teaches students how to write spreadsheet formulas with absolute and relative cell references. In the grounded feedback version of the tutor, Excel evaluates each formula that the student enters. From Excel's feedback (providing the calculated values for each formula), the student can determine if the original formula was correct. For example, one problem asks students to calculate the interest owed on a loan of $10,000, at various interest rates (Fig. 1.11). The problem content was designed so that students would likely be able to calculate each of the interests owed, or at least recognize clearly incorrect values (Fig. 1.11a). Fig. 1.11b shows the student-entered formula "=B2*A5" in the cell B5 (entry shown at the top), which multiplies the loan amount by the 1% interest rate. Excel responds with "$100" in B5 and this value matches the student's expectations. However, when the student copies the formula from B5 and pastes it in cells B6 through B8 (Fig. 1.11c), Excel's values do not match the student's expectations. The interest owed at the 5% rate is shown to be $0. Upon inspecting the formula for that erroneous cell (top of Fig. 1.11c), the student is intended to see that the interest rate is not being multiplied by the loan amount, but by the cell directly underneath. Excel multiplied the wrong cells because the student used a relative instead of an absolute reference (the correct formula for B5 is "=B$2*A5"). If the student cannot fix the error independently, the tutor provides step-by-step guidance. The grounded feedback tutor (which Mathan and Koedinger called an "Intelligent novice model spreadsheet tutor") was compared to a version that gave explicit interactive support as soon as students entered an incorrect formula. In this control condition, students had to generate the correct formula before pasting it into multiple cells. In both conditions, the tutor offered text hints if students needed them. In the intelligent novice condition, students could (1) see how Excel responded to incorrect formulas; and (2) try to recognize and correct their own errors before the tutor jumped in.

An experiment with 49 adult job seekers using a pretest-intervention-posttest design showed that, like the ANIMATE experiment, while both groups improved from pre- to posttest, the grounded feedback group improved more (Mathan & Koedinger, 2003, 2005). Students in the grounded feedback condition showed significantly better learning from pretest to posttest on all of their measures, with substantial effect sizes (across all treatment-to-control comparisons) for problem solving (effect size: .50),

conceptual understanding (effect size: .59), transfer (effect size .43), and retention (effect size: .33). These strong results show the additive benefit of grounded feedback in a learning environment that already provides text hints.
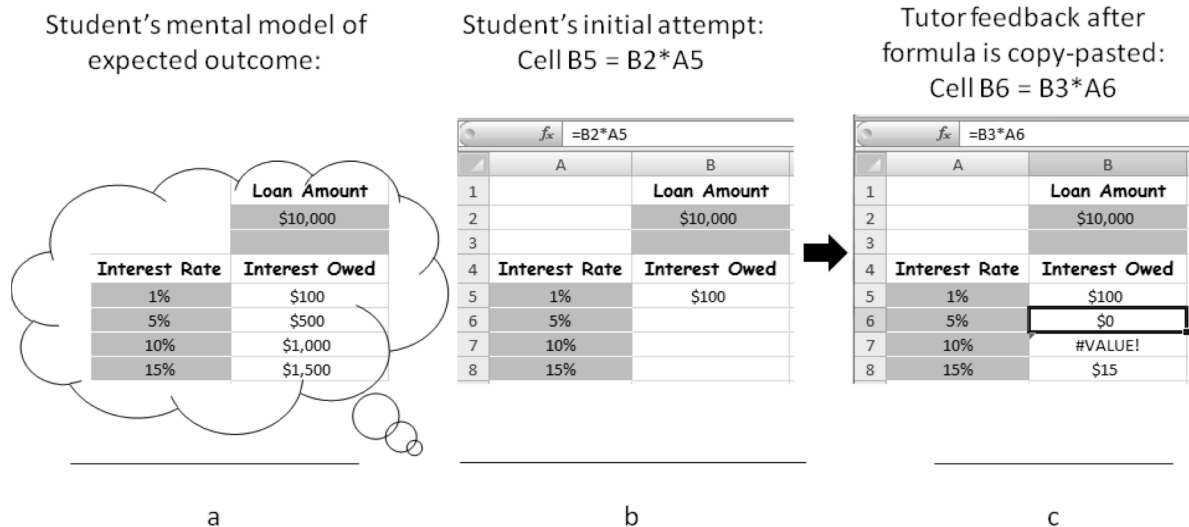

Fig. 1.11. a) The student mentally calculates the interest rates. b) The student multiplies the first interest rate (cell B2) by the loan amount (cell A5), yielding the correct interest for cell B5. c) When that formula is copied and pasted in cells B6-B8, it multiplies the interest rates by cells under the loan amount. This result does not match the student's expectations.

1.3.3 Invention: Grounded Feedback for Learning Principles

Though the previous examples of grounded feedback were intended to teach content immediately, grounded feedback can also prepare students for future learning. One example is an invention activity for finding formulas for variance and mean deviation (Schwartz & Martin, 2004). The expectation was not that students would be able to find the correct formulas on their own, but rather that the process of invention would help students find the features that such formulas would need to include, and would then provide a solid foundation for learning the correct formula from a teacher. In the invention activity, students were given a set of contrasting cases of points on graphs, which the students could rank intuitively by spread (Fig. 1.12). The students then generated formulas and could see if their invented formulas produced the same rankings. Here, the rankings are the grounded feedback. If students saw that their invented formulas produced rankings that differed from their initial intuition, they would know the formula was not correct. When students looked more carefully at the incorrectly ranked graphs, they were intended to see important differences between the graphs that the invented formulas would need to take into account (e.g., number of data points). While the comparison between a formula's expected and actual rankings allows students to check their own work, the carefully constructed contrasting cases allow for inferences about the formula's errors.

29

Schwartz and Martin (2004) compared two instructional sequences: the invention activity followed by a 5-10 minute lecture on the correct formula followed by practice with the correct formula vs. a lecture on the correct formula followed by practice (with the same total amount of time allotted for both sequences). Assessments measured not only students' target knowledge on the correct formulas, but also for their ability to learn from a worked example: The posttest included a far-transfer question that required knowledge of concepts presented in an embedded worked example. A pre-to-post test comparison of the two conditions found that while both groups learned about the same amount of target knowledge, the invention students were better able to learn from the worked example (Schwartz & Martin,

Here are four grids showing the results from four different pitching machines. The X represents the target and the black dots represent where different pitches landed. Your task is to invent a procedure for computing a reliability index for each of the pitching machines. There is no single way to do this, but you have to use the same procedure for each machine, so it is a fair comparison between the machines. Write your procedure and the index value you compute for each pitching machine using the grids below.
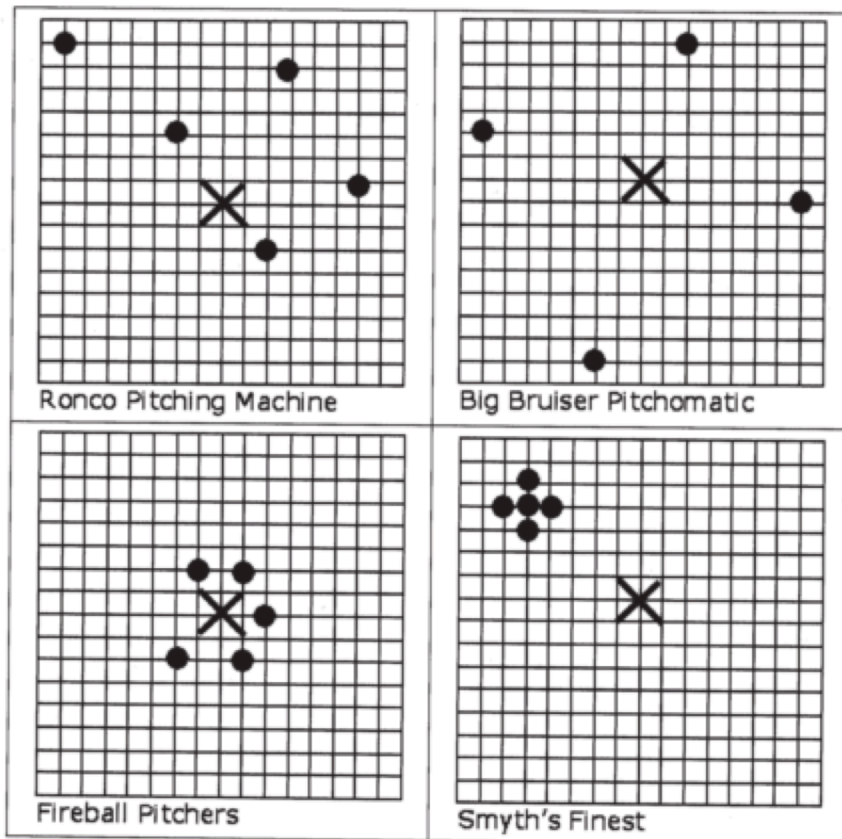

Fig. 1.12. An invention activity. The intuitive rankings of reliability provide grounded feedback.

2004). In other words, the invention activity helped prepare the students for a future learning opportunity.

Why didn't Schwartz and Martin's study show a learning benefit for target knowledge for the invention group, which learned with grounded feedback, when the ANIMATE and Excel studies did show such a benefit? The key difference is the subject of the grounded feedback. With ANIMATE and Excel, the purpose of the instruction was for students to generate correct systems of equations and formulas (respectively), and the two systems gave students grounded-feedback practice with those target skills. In the invention activity, the goal was not to have students practice the correct use of the formula. Instead, the goal was to help students explore how the inclusion of different features affected their invented formulas, and that was the subject of the grounded feedback. Both groups practiced using the correct formula in the same way, with the same kind of feedback, which explains why the invention/grounded group had similar scores on target posttest items as the control group. However, the invention group, which got grounded feedback on the principles, did demonstrate better understanding of those domain principles, as indicated by their ability to learn from an embedded worked example on the posttest. Since students working with invention activities almost never come up with the canonical solutions, such activities fall under the umbrella of "productive failure" (Kapur, 2009). We hypothesize that grounded feedback is a necessary element for making failure productive. For example, if Schwartz and Martin's experiment had used the charts in Fig. 1.13, students would have had much more difficulty producing intuitive rankings for the charts. When comparing their calculated rankings to the intuitive ones, students would not necessarily trust their intuitive rankings more. Note that Fig. 1.13 preserves the four contrasting cases: B shows the most reliable machine, but points are not centered around the X; A shows points clustered around the X; C shows a cluster of four and an outlier; D has four points instead of five, is clustered around the X, and is least reliable. Yet without the grounded feedback provided by the intuitive rankings, students would likely flounder unproductively. However, as with previous hypothesized comparisons in this review, I am not aware of experiments that compared grounded and ungrounded forms of invention activities.
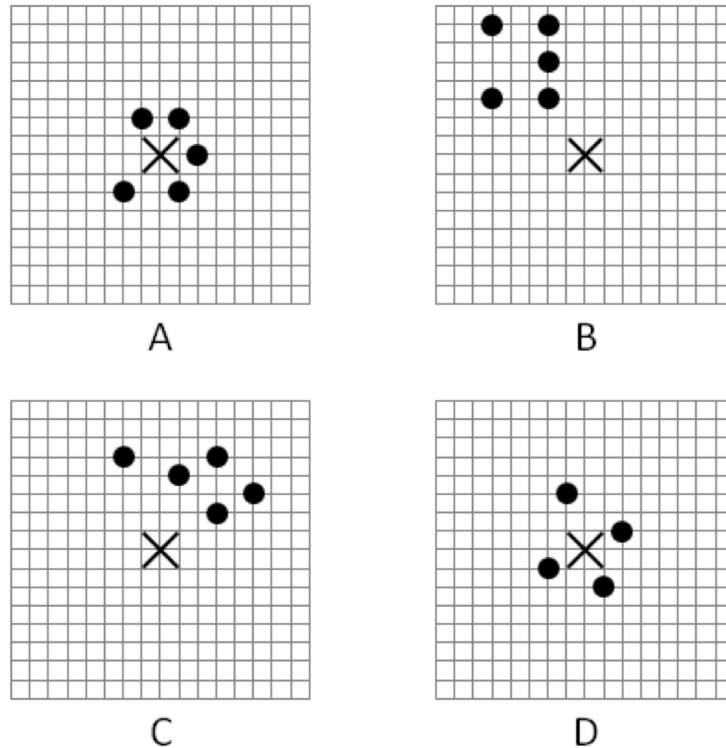
Fig. 1.13. Charts with non-intuitive rankings of mean deviation do not provide grounded feedback.

1.3.4 Experiments with Inconclusive Results or No Differences in Learning

While the experiments discussed above showed robust learning benefits for grounded feedback, other related work has found no differences in learning or inconclusive results. One study on fractions found that linked representations did not outperform worked examples (in the context of a larger investigation of learning with multiple representations; Rau, Aleven, Rummel, & Rohrbach, 2012). The linking was similar to grounded feedback in that students worked with number lines (a less-familiar representation) and got feedback on their actions with fraction rectangles or circles. However, the linking was not a consistent implementation of grounded feedback: in some cases the second representation was static and only depicted the correct answer, and in some cases the dynamic representations could only reflect a subset of students' inputs on the number line. This study suggests that worked examples may be more effective than grounded feedback when three representations are involved, but due to the inconsistent implementation of grounded feedback in this system the results remain inconclusive.

Another study, on algebraic transformations, compared four conditions: grounded feedback, problem-level right/wrong feedback, problem-level right/wrong

feedback with on-demand demonstration of transformation steps, and no feedback (Yerushalmy, 1991). In each version of the tutor, students were given an algebraic expression and had to transform it into a different format (e.g., given x(x-2)3+7(x-2), change it into the format Ax^2+Bx+C). In the grounded feedback condition, students were shown three graphs: one of the original expression, one of the student's current work, and one showing the difference between them. An experiment with 7th graders used a pretest-intervention-posttest design to measure learning outside the tutor; test data was reported for 17 students. All groups improved from pre- to posttest on the target content, without significant differences in learning between the conditions. While this study is quite small and likely underpowered, Yerushalmy identified some qualitative pros and cons of the grounded feedback. First, it appeared that the feedback was indeed useful in helping students evaluate if their attempts were correct – compared to the no-feedback condition, grounded feedback students performed more steps per problem, did more relevant debugging, left fewer uncorrected errors, and ultimately solved more problems correctly during tutoring. For students with high prior knowledge, the graph feedback helped them locate which term was the source of their error. However, other students, especially those with low prior knowledge, reacted to the grounded feedback as problem level right/wrong feedback or simply tried to eliminate the difference graph without looking for the source of the error, a form of gaming the system.

1.4   Conclusions

This chapter presents *grounded feedback,* defined by four criteria: 1) The feedback is intrinsic to the domain and semantically equivalent to the student's inputs; 2) Students can easily envision a correct goal state for the feedback; 3) The input format matches the domain learning goals; and 4) The feedback affords inferences on errors. Prior work provides experimental support for grounded feedback (Nathan's ANIMATE, 1998; Mathan & Koedinger's Excel tutor, 2003), but also shows that grounded feedback is not simple to implement (Rau et al., 2012). From a theoretical perspective, grounded feedback is likely to lead to more robust learning than related forms of feedback, but sufficient empirical comparisons have not yet been conducted. Overall, I believe there is enough promising support for grounded feedback to warrant further investigation. In particular, future work should examine if each feature of grounded feedback is important; how grounded feedback and explicit supports are best combined; and how students interact with grounded feedback. I hypothesize that students learning with grounded feedback will engage in sense making through mapping: from the input representation to the feedback representation, and from the target knowledge to their prior knowledge. In this manner, grounded feedback would provide students practice in building knowledge through inference and in checking their work using their own prior knowledge. If grounded feedback does indeed strengthen these skills, over time it may help students learn both domain knowledge and the metacognitive skills necessary to become more reflective learners.

2   Grounded Feedback for a Fraction Addition Tutor

Summary. Standard intelligent tutoring systems give immediate feedback on whether students' answers are correct. This prevents unproductive floundering, but may also prevent the development of useful error detection and correction skills. This chapter presents the first iteration of the grounded feedback tutor for fraction addition. In two think-aloud studies with 6 and 5 fifth-graders, students were able to solve more fraction addition problems with the tutor than with paper and pencil. Further, students were able to correctly interpret the feedback, and used it to find and fix their mistakes – without correctness feedback.

2.1   Initial Tutor Design

In the fraction addition tutor, the grounded feedback takes the form of rectangular fraction bars. For each proper fraction n/d, a fraction bar is divided into d parts, with n colored in. The tutor shows fraction bars representing the two addends, and shows the sum as a combination of those magnitudes. As the students enter the converted and sum fractions, the tutor reflects those quantities in the fraction bars (Fig. 2.1). The rectangles are intended to allow for easy comparison between the given fractions in the problem and student-generated converted and sum fractions. This example-tracing tutor was built with the Cognitive Tutor Authoring Tools (CTAT; Aleven, Mclaren, Sewall, & Koedinger, 2009).

The fraction bars are intended to show meaningful consequences of students' actions that make visible the intermediate and goal states of the problem solution, without giving away the answer. The design goal is that visual feedback will allow students to see if original addends and converted fractions are equivalent, and whether their answer fraction is equivalent to the sum of the two given fractions. The tutor does not give explicit feedback on the correctness of intermediate steps during problem solving.
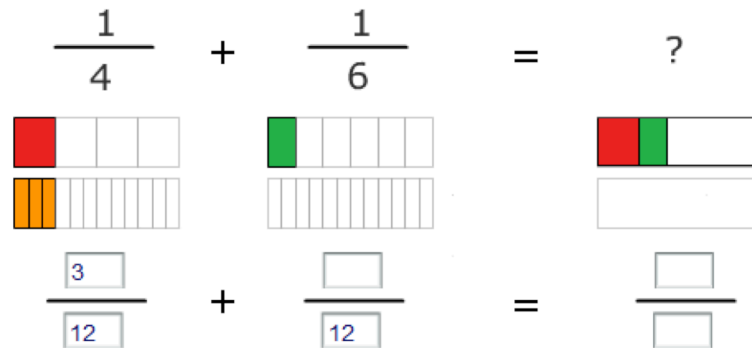
$$\frac{1}{4} \quad + \quad \frac{1}{6} \quad = \quad ?$$

$$\frac{3}{12} \quad + \quad \frac{\square}{12} \quad = \quad \frac{\square}{\square}$$

Fig. 2.1. The grounded feedback tutor interface. The fraction bars in the first row are given, and include both addends and the multi-colored sum fraction. The bars in the second row update based on the student's entries in the bottom row.

2.2   Study 1a: Think-Aloud

In a think aloud study, participants are asked to perform a task while verbalizing their thoughts (Gomoll, 1990), a useful technique for designing tutoring systems (Lovett, 1998). The first think aloud assessed whether the fraction bar feedback met the student-based criterion for grounding: if students could use the feedback to detect errors, and if that detection was easier with the feedback than with the numeric symbols alone. Further, the think-aloud explored what prior knowledge the grounded feedback elicited, and if students would integrate a representation of magnitude with their procedural steps for solving a fraction addition problem. With a collaborator, I conducted the initial think aloud, with paper-and-pencil problems followed by tutor problems.

2.2.1   Participants, Materials, and Procedure

Six fifth graders from an all-girls school in Pittsburgh volunteered to participate in the think alouds at their school during the school day. According to their math teacher, the girls had learned about fractions but not fraction addition. Each student participated individually in a 20-25 minute think aloud session with the experimenters. Students were asked to solve fraction addition problems with pencil

and paper, explaining their thoughts and steps out loud. Next, they were asked to explore the fraction addition interface and explain out loud what they understood to be happening. Then they solved up to three tutor problems. At times, if students were stuck the experimenters would provide verbal hints, though the experimenters did not evaluate students' work. Students solved the following tutor problems in order: 2/8 + 3/8 (same denominator), 1/3 + 2/9 (one denominator is a multiple of the other), and 1/4 + 1/6 (unrelated denominators). Paper problems were also given from those categories in that order, though problems were not all the same. This think aloud was intended to get an initial sense for how students interacted with the tutor. The inconsistencies with experimenters' hints and differences in paper problems were resolved in the next think aloud.

2.2.2   Results and Discussion

Students appeared to understand that the fraction bars reflected the quantities that the students entered, and when the colored areas of two fractions were equal, the fractions were equivalent. Table 2.1 shows the percentage of students who correctly solved each problem without verbal hints from the experimenters.

Problem Type	Paper	Tutor
Same denominator	4 (66%)	6 (100%)
One multiple of other	2 (50%)	3 (50%)
Unrelated denominators	0 (0%)	1 (20%)
Total	6/13 (46%)	10/17 (58%)

Table 2.1 Number of participants who correctly solved each problem without verbal hints from the experimenters.

The problem categories were progressively harder for students. On paper, four out of six students correctly solved the same denominator problem, two out of four solved the one-multiple-of-other problem and none out of three solved the unrelated denominators problem. With the visual feedback tutor, all six students solved the same denominator problem, three out of six solved the one-multiple-of-other problem, and one out of five solved the unrelated denominator problem. When students got stuck with the tutor, the experimenters gave them verbal hints. With hints, all students correctly solved the one-multiple-of-other problem and four out of five correctly solved the unrelated denominator problem.

Students used the fraction bars to evaluate their initial attempts and to generate answers. One student solving 2/8 + 3/8 initially added the numerators and denominators, yielding 5/16. The grounded feedback produced a fraction bar for 5/16 that was smaller than the multi-colored bar representing the target sum. The student noticed and she correctly changed her answer to 5/8. One student who wanted to convert 1/3 to ninths entered nine as the denominator, and then counted the divisions to see how many ninths equaled one third. She explained that she found 3/9 both from the pictures and because three times three was 9 (Fig. 2.2).

Students looked surprised when the visual feedback showed that their fractions were smaller or larger than the target fractions, indicating that the consequences of the students' actions contradicted their expectations. Students made, recognized, and corrected errors on their own. Students used the visual feedback both to evaluate fraction equivalence and produce numerators. If the colored areas of two fraction bars were the same, the students took the two fractions being represented as equivalent. This caused one misconception when a student thought 3/7 was equal to the sum of
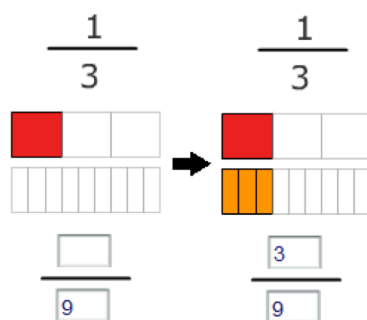

Fig. 2.2 Using visual feedback to convert 1/3 to 3/9

1/4 and 1/6. The difference between those two amounts is so small that the fractions bars appeared equal. Students also used the visual feedback to test possible denominators by generating images with two consecutive numerators. For example, one student tried 12 as a denominator for the sum of 1/3 + 2/9. She found that 6/12 was too small but 7/12 was too big, so 12 could not be the appropriate denominator for the answer.

Students' behaviors indicated that the feedback was grounded: students expected their converted fractions to be equivalent to the original addends, and expected their sum fractions to be equal to the combined magnitudes of the two addends. They noticed when corresponding fraction bars did not align, and interpreted that state as an error. The strategies that students used with the grounded feedback would not have been possible without it. It is possible that engaging with the visual feedback allows students to explore properties of fractions that are not self-evident from pencil and paper alone. However, the fraction bar feedback could encourage students to guess until the bars look the same, a form of gaming the system. Finally, the visual feedback can be misleading when differences between two fractions are very small, which could harm students' learning.

Students correctly solved more problems with the tutor than they did with pencil and paper. However, the fraction bars alone not sufficient for guiding all students. These students benefitted from verbal hints from the experimenters, and overall students correctly solved 17/18 tutor problems. Further, two students who completed all three tutor problems returned to their paper problems and corrected their mistakes without additional hints. These successes are consistent with the hypothesis that these students developed better understanding of fraction addition from the grounded feedback.

Although the grounded feedback was useful for some problem steps, it was not sufficient for students to solve all of the problems. Students did not get any support in choosing a next step or picking a denominator. Students got stuck when they did not know to convert the given fractions to ones with a common denominator. Students who knew they were supposed to convert the fractions also got stuck because there was no support from the interface to tell them which denominator to try. The next iteration of the tutor aimed to address these issues.

2.3   Study 1b: Revised Tutor and Think-Aloud

Since students floundered when they did not know to find a common denominator and when they did not know which denominator to try, the second version of the tutor included a 3-level succession of on-demand text hints that first told students to find a common denominator and then gave a general and then problem-specific suggestion for how to do so. The hints did not tell students the answer.

2.3.1   Participants, Materials, and Procedure

Five students from the first study volunteered to participate in the second (the last student was sick). Students solved one problem from each category on paper (1/9 + 4/9, 2/3 + 1/6, and 1/2 + 1/5) and with the tutor (1/7 + 2/7, 1/4 + 3/8, and 2/5 + 1/3). One student did not have time for the last tutor problem.

2.3.2   Results and Discussion

For each problem category, the five students correctly solved more problems with the tutor than on paper (Table 2.2). One student did not start the unrelated denominators tutor problem.

Problem Type	Paper	Tutor First Attempt	Tutor
Same denominator	4 (80%)	4 (80%)	5 (100%)
One multiple of other	3 (60%)	1 (20%)	4 (80%)
Unrelated denominators	1 (20%)	1 (25%)	3 (75%)
Total	8/15 (53%)	6/14 (43%)	12/14 (86%)

Table 2.2 Problems solved correctly without text hints.

As in the first think-aloud, the problem categories were successively more difficult: on paper, four students correctly solved the same denominator problem and only one correctly solved the unrelated denominators problem. Students' first attempts with the tutor reflect their problem solving without the grounded feedback. Students' first attempts with the tutor were no more successful than their work with paper, suggesting the tutor problems were at least as difficult as the paper problems, and students did not do better with the tutor because it came after the paper. With the

grounded feedback alone, all five students solved the same denominator problem, four solved the one multiple of other problem, and three of the four solved the unrelated denominator problem correctly. With the text hints built into the tutor, students solved all attempted tutor problems.

Students' comments during the think aloud showed how they connected the tutor's grounded feedback to their prior knowledge. For example, one student converted 1/4 to 1/8, but then changed it to 2/8 after seeing the fraction bar. The student explained, "a) I looked at the picture and realized they weren't matched up and b) I realized that I'd doubled the bottom but not the top." The interface already displayed the given fraction 1/4, and the student saw that the fraction she had entered, 1/8, was much smaller than 1/4. The difference between her expectation (that the pictures should match) and the consequences of her action (that they did not match) prompted her to review her procedure and check for errors.

Another student's comment illustrates the benefit of the built-in verbal hints and how students used the grounded feedback for evaluation. One student got stuck on the unrelated denominator problem, and the experimenter told her to ask the tutor for a hint. The student read two hint levels, which first told her to find a common denominator and then to think about multiples of the two given denominators. The student tried 15 as a common multiple of three and five, explaining, "I put 15 here to see how many fifteenths. I think five fifteenths going to for one third. I'm not sure." After seeing the feedback, she exclaimed "Oh yeah! I was right." The visual feedback convinced her that she was correct. The student then used this strategy to go back and correct the corresponding paper problem.

Again, this think aloud found good evidence of the feedback being grounded for these students, including recognizing errors from the visual display. Further, students seemed to engage in productive sense making and error-correction. Students demonstrated some positive use of the next-step hints available in this version, but also non-use in cases where they were clearly stuck.

2.4   Limitations and Conclusion

Students' interactions with the fraction bar feedback indicate that the student-based criterion for grounding is met: students were able to use the fraction bars to evaluate their own work. The grounded feedback tapped students' prior knowledge of the role of magnitude in the fraction addition procedure. Specifically, students knew that their converted fractions should be equivalent to the original addends, and their sum fractions should equal the combined magnitude of the addends. The grounded feedback alerted students to most magnitude discrepancies, and students interpreted those discrepancies to mean that their work was incorrect. However, these participants came from an academically rigorous private school and are likely not a representative sample of 5th graders in general. Further, although I used three paper problems as a crude pre-test, since I did not assess any other prior knowledge I do not know what other prior knowledge is needed for students to successfully interpret the feedback. Finally, because of the nature of a think-aloud study, students worked with

the tutor while continuously giving self-explanations, which may have made the tutor appear more effective. This study did not examine learning effects, though students did self-correct while using the tutor and there were six instances of students correcting their errors on the paper problems after working with the tutor (out of 14 incorrect paper problems). These outcomes suggest that a grounded feedback tutor can help students learn, a hypothesis that will be tested experimentally in the next chapter.

3 Comparing Grounded and Correctness Feedback in Fraction Addition Tutors

Summary. 128 fifth graders completed an experiment comparing two types of feedback in an intelligent tutoring system for fraction-addition. *Correctness feedback* indicated when each step was right or wrong, and *grounded feedback* showed fraction bars as conceptual scaffolds, requiring more student interpretation. Correctness students solved the tutor problems more efficiently. Grounded feedback students improved more than correctness students overall between the pre-test and delayed-test, suggesting that grounded feedback may be more beneficial for long-term learning.

3.1 Grounded Feedback Tutor Design

The grounded feedback tutor for this study was based on the initial design presented in Chapter 2. New interface elements were added to support problem-planning and fraction conversion. Upon starting a problem, the tutor shows the fractions in the problem statement, with a question mark representing the sum. Under each addend, a checkbox states, "I need to convert this fraction," with "I'm ready to add" under the sum (Fig. 3.1). When the student selects a fraction to convert, a conversion interface opens, with input areas for the intermediate multiplication step (Fig. 3.2). Creating separate interface elements for fraction multiplication is intended to make this step more explicit for students, and to allow students actions with this step to be captured

by the tutor logs. When the student checks the box "I'm ready to add," the tutor copies the converted fractions to the addition area. If the student has not converted one of the fractions, the original addend is copied to the addition area (Fig. 3.3).

To prevent unproductive floundering, the tutor incorporates three levels of on-demand text hints, with the last hint giving the answer for the current step. A three-

Fig. 3.1 Upon starting a problem, the student sees the problem statements and prompts for planning the next problem-solving step.

level hint sequence is given based on which steps the student has already completed, starting with deciding what initial action to take (convert if the fractions have different denominators, add if the denominators are the same). If conversion is necessary, the hints start with identifying a common denominator, finding the numbers to multiply by, and then finding the numerator for the converted fraction.

The model of desired performance with the grounded feedback tutor is based on engagement with "subjective" errors (Ohlsson 1996), also termed "intelligent novice" (Mathan & Koedinger 2005). That is, students are permitted to make mistakes in the hope that students will recognize when the consequences of those actions violate their expectations. For example, if a student started the problem 2/4 + 1/8 by checking the "I'm ready to add" button and entering 3/12, the grounded feedback would show that the student's answer is much smaller than the target sum. While this action is objectively incorrect, it may be a good action to take from the perspective of engaging
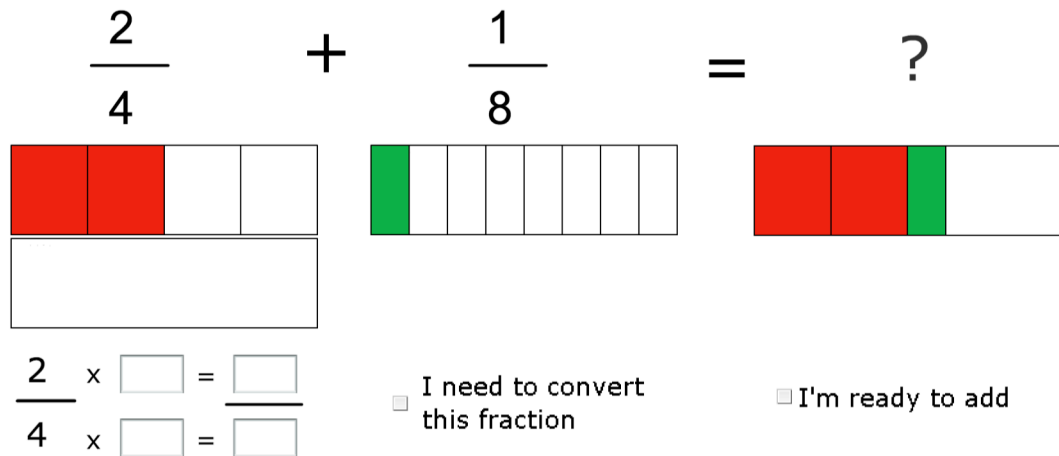
$$\frac{2}{4} \quad + \quad \frac{1}{8} \quad = \quad ?$$

$$\frac{2}{4} \quad \times \boxed{} = \boxed{}$$
$$\quad \times \boxed{} = \boxed{}$$

☐ I need to convert this fraction   ☐ I'm ready to add

Fig. 3.2 After selecting "I need to convert this fraction" for 2/4, the converting interface opens. The original addend is copied at the left, with input areas for multiplication next to it. The converted fraction, to the right of the equal signs, drives the fraction bar immediately above.
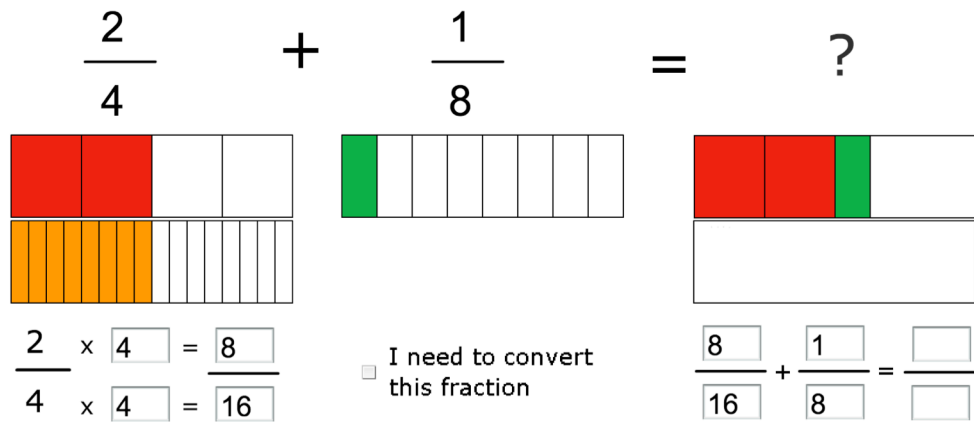
$$\frac{2}{4} \quad + \quad \frac{1}{8} \quad = \quad ?$$

$$\frac{2}{4} \quad \times \boxed{4} = \boxed{8}$$
$$\quad \times \boxed{4} = \boxed{16}$$

☐ I need to convert this fraction

$$\frac{8}{16} + \frac{1}{8} = \frac{\boxed{}}{\boxed{}}$$

Fig. 3.3 After converting 2/4 to 8/16 and selecting "I am ready to add," the addition interface opens, copying the converted fraction 8/16. Since 1/8 was not converted, the original addend is copied. The sum fraction, to the right of the equal sign in the addition area, drives the fraction bar immediately above it.

the student as an active learner: seeing the consequence of the mistake and recognizing the action as incorrect may be a more powerful learning opportunity than simply being prevented from going down that path in the first place. However, if a student does not expect their sum to have the same magnitude as the multi-colored fraction bar, a mismatch will not be recognized as an error. Conversely, a student may

solve a problem inefficiently, such that the magnitude constraints are not violated (e.g., converting both 2/4 and 1/8 to sixteenths). In that case, a student might miss the learning goal of that problem (e.g., to illustrate that when one denominator is a multiple of the other, only one fraction needs to be converted). Therefore, while correctness feedback is not provided for any intermediate problem steps, problem-level correctness feedback is provided when the student presses the "done" button. If the problem has been solved correctly, the student moves on to the next problem. If the student made a mistake, or solved the problem inefficiently, a message appears in the hint window tell the student that they are not done yet, and suggesting they ask for a hint if they're not sure what to do. To solve a problem efficiently, students must use the given denominator for addends with the same denominator; the larger denominator when one is a multiple of the other; and the least common multiple or the product when the two denominators are unrelated.

3.2   Correctness Feedback Tutor Design

The correctness feedback tutor uses the same basic interface as the grounded feedback tutor, but without the fraction bars. Students see the same checkboxes for converting and adding when the problem starts. However, this tutor provides immediate correctness feedback on each step, coloring inputs green if correct, and red otherwise. The correctness feedback tutor follows the "objective" (Ohlsson, 1996) model of errors, also called the "expert model" (Mathan & Koedinger 2005). That is, if a student's action is not on an efficient solution path, it is marked as incorrect. At the problem-planning stage, incorrect paths are closed off. The tutor only permit students to open the conversion and addition interfaces if those actions are objectively correct given the current problem state: students may only convert fractions when the addends have different denominators, and only the fraction with the smaller denominator when one denominator is a multiple of the other. Students may only open the addition interface when starting a same-denominator problem, or after the necessary fractions have been converted correctly. Showing the same inputs as Fig. 3.3, the correctness feedback tutor marks all of the converting inputs as incorrect, and does not allow the student to add, since the student has not yet converted the first fraction to eighths (Fig. 3.4). Unlike the grounded feedback tutor, the correctness tutor does not allow students to erase correct inputs. Like the grounded feedback tutor, the correctness tutor ensures that students have solved the problem correctly before moving on to the next problem. The correctness tutor also offers three levels of on-demand texts hints for each problem step.

Fig. 3.4 The correctness tutor marks the conversion of 2/4 to 8/16 as incorrect, since the efficient denominator is 8. The tutor will not permit the student to open the addition area until 2/4 is correctly converted to eighths.

3.3   Study 2: Comparing Correctness and Grounding

This study uses paper assessments to measure learning outside the tutor, and compares the grounded feedback tutor to a robust control – correctness feedback. The target student for this study can convert fractions but is not yet fluent with the addition of proper fractions with sums below one. This study also examined the effect of introducing the fraction bars by first relating them to another concrete representation.

3.3.1   Materials: Tutor Introduction

Students are introduced to the tutor with simple arithmetic problems to practice using the tutor interface. Students are asked to explicitly acknowledge that the tutor is intended to help them learn and practice, not test them. Students are also encouraged to ask for hints if they get stuck (Fig. 3.5).
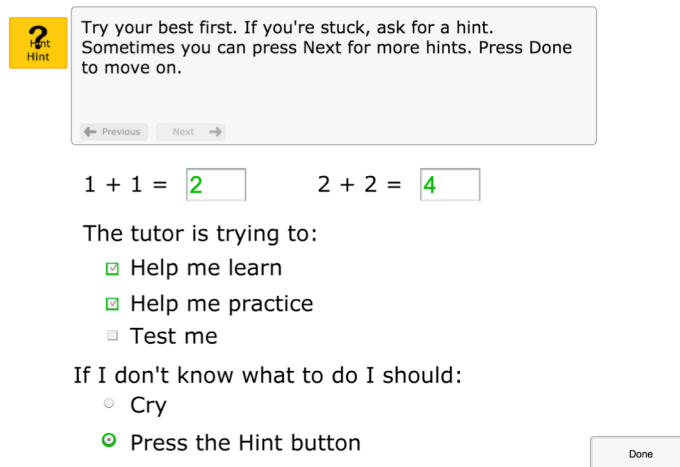
Fig. 3.5 The tutor introduces students to the input elements (text input, checkboxes, and radio buttons) before teaching new content. Students are encouraged to view the tutor as a learning opportunity.

3.3.2 Materials: Addition Instruction

After the introduction to the tutor, students are given brief instruction on naming fractions and fraction addition with like and unlike denominators, with immediate correctness feedback and on-demand text hints. This instruction is presented in the context of measurement, to provide a concrete representation that is different from the fraction bars. That way, the addition instruction will not give students practice with the grounded feedback. The first instruction page asks students for the length of two bugs sitting on rulers with each inch divided in thirds. Then students are asked to find the combined length of both bugs (Fig. 3.6). In the next instruction page, students name the lengths of two bugs sitting on rulers with the same unit size, one where the unit is divided in two pieces and the other in three pieces. Students are shown the combined length of both bugs first on one ruler and then the other, to illustrate that the numerators cannot be added when the denominators are not equal (Fig. 3.7). Students are guided through converting both fractions to a common denominator (Fig. 3.8) and then adding (Fig. 3.9).

The answer says that when each inch is divided into 3 equal pieces, both bugs together are 5 pieces long. Look at the ruler and then press the Done button to move on.

← Previous   Next →

$$\frac{4}{3} + \frac{1}{3} = \frac{5}{3}$$
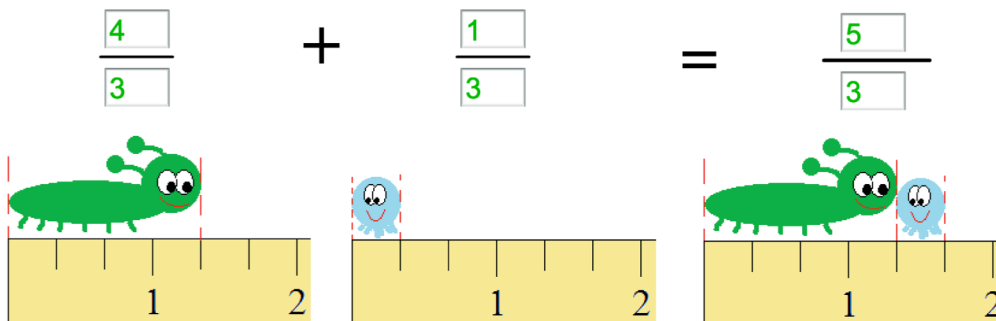
Done

Fig. 3.6 After naming the lengths of each bug, the student adds to find the total length. Interface elements appear in sequence as the student moves through the problem steps, to prevent the student from being overwhelmed.

We can't add the top numbers when the bottom numbers are different. Lets see why. Is the answer two thirds? Choose an answer next to the picture.
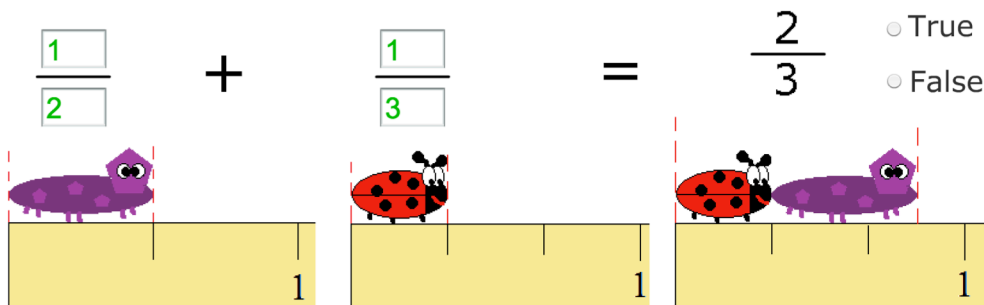
← Previous   Next →

$$\frac{1}{2} + \frac{1}{3} = \frac{2}{3}$$

○ True

○ False

Fig. 3.7 After naming the lengths of each bug, the student is shown the combined length of both bugs, first on a ruler divided in thirds and then on a ruler divided in halves, to illustrate that one cannot add the numerators when the denominators are not the same.
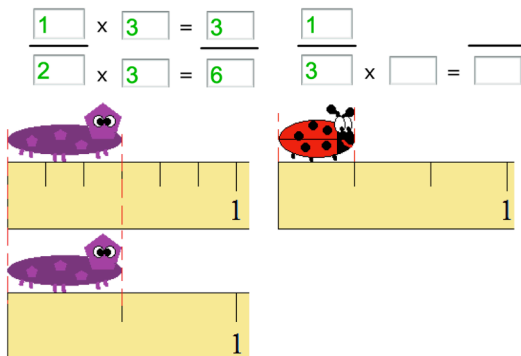
47

Fig. 3.8 The student is guided through converting both fractions to a common denominator. The illustrations show each bug being measured by two different rulers, corresponding to the original addend and the converted fraction.
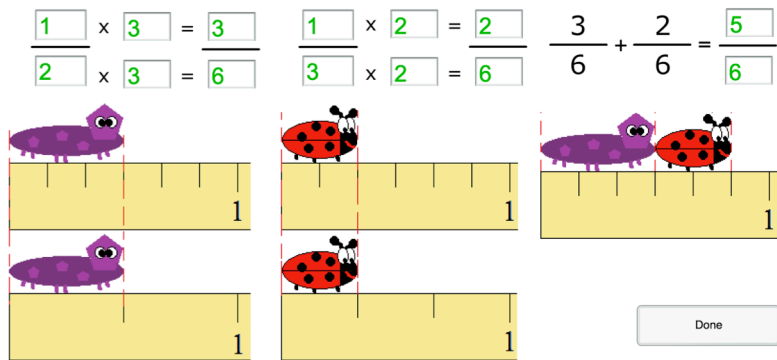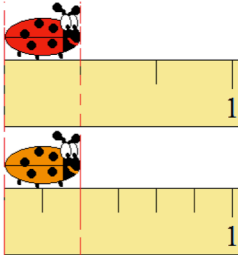

Fig. 3.9 After converting the fractions, the student adds them to find the combined length of both bugs. After finding the sum, the addition is illustrated with both bugs on a ruler where the unit is divided in sixths.

3.3.3 Materials: Transition from Bugs to Rectangles

After the addition instruction, students in one condition are given guidance on relating the concrete bug and ruler representation to the more abstract fraction bar representation (Fig. 3.10). The multi-colored sum fraction is related to the combined length of both bugs (Fig. 3.11). Students in this condition are also encouraged to convert fractions by multiplying instead of by counting the fraction bar pieces after finding a denominator.
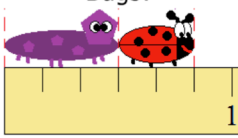
Fig. 3.10 Transitional instruction relates the bug representation to the fraction bar representation in the context of equivalence.

Fig. 3.11 Transitional instruction relates the bug representation to the multi-colored sum fraction bar.

3.3.4   Participants and Method

Six fifth grade classes from a public school near Pittsburgh elected to participate in the study (138 students participated in the first study day). Two classes each were high, average, and low achieving. The study took place in school's computer lab, during the normal school day. Over one double period (80 minutes), students completed a 15-minute pretest, worked with a randomly assigned tutor for up to 40 minutes, and then took a 15-minute post-test. Two weeks later, students took a delayed post-test. The three assessment forms were matched and counter-balanced. The paper-and-pencil tests included four fraction addition items, three evaluation items, three prior knowledge items, and three conceptual items based on released items from standardized tests (NAEP, PSSA, and MCAS). Students completed the addition section before starting the other questions to ensure that performance on the addition items did not reflect learning from other test items. Additionally, the pre-test included a reading comprehension item based on the tutor's text hints. However, some test forms inadvertently did not include the reading comprehension item, so it is not included in the analyses below. The three test forms were counter-balanced between assessment times as follows: 16 students completed the tests with an ABC order; 21 with ACB, 10 with BAC, 16 with BCA, 21 with CAB, and 20 with CBA. 15 students were inadvertently given the same test form either at pre-test and post-test or post-test and delayed-test. 9 students were inadvertently given questions from two test forms during at least one test time (e.g., the addition items from form A paired with the non-addition items from form B). In analyses for overall scores that include test form, the cases with questions from more than one test form are coded as a separate category (mixed).

This study included three tutoring conditions: Correctness, Grounded, and Grounded with Guidance. All tutors started with the tutor introduction, followed by the addition instruction. The Grounded with Guidance condition included the transitional instruction before the fraction addition problems. Each tutor included the same 20 fraction addition problems: 5 where the denominators were the same, 5 where one denominator was a multiple of the other, and 10 with unrelated denominators. Before starting the regular tutor problems, students saw the problem 1/2 + 1/3, to ease them into the tutor interface, since they had been guided through solving that problem with the bugs. Within each tutor condition, students were randomly assigned to one of three subsequent problem sequences of six problems each: one with two same-denominator problems, followed by two one-multiple-of-the-other problems, and then two unrelated-denominator problems; one with two same-denominator problems, followed by two unrelated-denominator problems, and then two one-multiple-of-the-other problems; and a randomly-determined sequence: same, one-multiple-of-the-other, unrelated, unrelated, one-multiple-of-the-other, same. The subsequent problem sequence was determined randomly.

3.3.5 Hypotheses

1) The brief instruction in the Grounded with Guidance condition will improve students' learning compared to the Grounded condition.
2) Compared to the Correctness condition, both Grounded conditions will result in greater overall learning.
3) Students in the two grounded conditions will understand the grounded feedback.
4) Initial question sequence will affect students' learning.

3.3.6 Overall Results
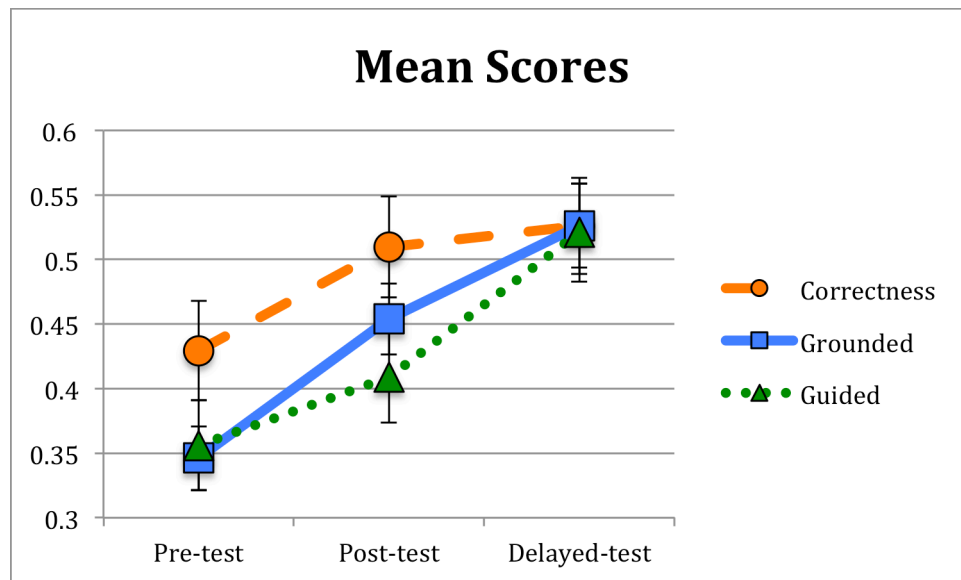
Fig. 3.11 Mean scores on each overall assessment, by condition, per test time, with bars showing standard error of the mean.

Of the 138 students who participated in the first study day, 9 students did not take the delayed post-test, and the anonymous ID matching of one student was lost between the first study day and the delayed test (although the student participated on both days). Attrition from each study condition was 8.3% for Correctness (4/47), 1.9% for Grounded (1/51) and 12.8% for Grounded with Guidance (5/40). Fisher's exact test on the number of students who completed all versus some of the study for each condition indicates that the differences between conditions are not significant (p=.13). The results that follow are based on the 128 students who completed all parts of the study (45, 48, and 35 students in the Correctness, Grounded, and Guided conditions, respectively). The reading comprehension item was inadvertently left off of 6 test forms, so that item is not included in the analysis. The PFL items are also not included in these analyses, as there were no matched questions for those items at the pre and post test. Figure 3.11 shows the mean scores for each assessment time, by condition, and Figure 3.12 shows the mean scores for the addition items. Table 3.1 shows process measures: mean number of tutor problems started, mean number of

hints per regular tutor problem, and mean number of seconds spent on the pre-instruction (tutor introduction, addition instruction, guided transition, and repetition of the problem ½ + 1/3). Two students each in the Grounded and Guided conditions did not start any regular tutor problems.
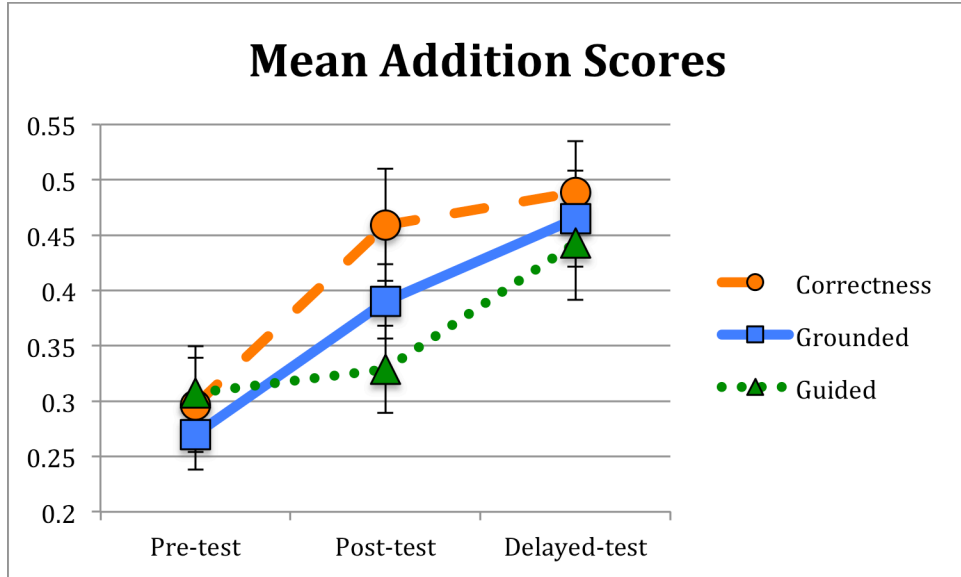
Fig. 3.12 Mean scores on the addition items, by condition, per test time, with bars showing standard error of the mean.

Condition	Pre-instruction duration (min:sec)	Regular tutor problems attempted	Hints per regular tutor problem
Correctness	10:08 (0:29)	17.7 (.51)	2.4 (.47)
Grounded	18:34 (1:19)	10.7 (1.0)	8.5 (1.4)
Guided	22:45 (1:47)	7.3 (1.1)	7.3 (1.6)

Table 3.1 Mean number of seconds spent on the pre-instruction (tutor introduction, addition instruction, guided transition, and repetition of the problem ½ + 1/3), mean number of tutor problems started, and mean number of hints per regular tutor problem, with standard error of the mean in parentheses.

3.3.7 Results and Analysis for Hypothesis 1

To determine if the additional guided instruction caused a difference in students' progression through the tutor, I ran ANCOVAs on the number of regular tutor problems students started, seconds per tutor problem, hints requested per regular tutor problem, and duration of the total pre-instruction (tutor introduction, addition instruction, guided transition, and repetition of the problem ½ + 1/3). The ANCOVAs

included pre-test score as a covariate, with class achievement level and condition as fixed factors, and a condition by class achievement interaction term.

For the number of regular tutor problems attempted, the interaction was not significant (p>.2) so the model was re-run without it. In both models (with and without the interaction term) pre-test score was significant at p<.005, class achievement level was significant at p<.04 and condition was significant at p<.01. Estimated marginal means for the model without the interaction term show the Grounded condition solving more regular tutor problems than the Guided condition (10.87 vs. 7.25, evaluated with pre-test score at .349).

For the number of seconds per regular tutor problems attempted, the interaction was not significant (p>.3) so the model was re-run without it. In both models (with and without the interaction term) pre-test score was significant (p<.016), class achievement level was not significant (p>.4) and condition was not significant (p>.1). Estimated marginal means for the model without the interaction term show the Grounded condition taking 37 fewer seconds per problem than the Guided condition (2 minutes 30 seconds vs. 3 minutes 7 seconds, evaluated with pre-test score at .354). This model was run with 46 students in the Grounded condition and 33 students in the Guided condition, since two students in each condition did not complete any regular tutor problems.

For the number of hints requested per regular tutor problem attempted, the interaction was not significant (p>.2) so the model was re-run without it. In the model with the interaction term, pre-test score was significant at p=.033; without it, pre-test score was marginal at p=.06. In both models (with and without the interaction term), condition was not significant at p>.5. With the interaction term, achievement level was marginal at p=.06; without it, achievement level was significant at p=.038. Estimated marginal means for the main-effects model show the Grounded condition requesting slightly more hints per regular tutor problems than the Guided condition (8.6 vs. 7.7, evaluated with pre-test score at .355). This model was run with 46 students in the Grounded condition and 33 students in the Guided condition, since two students in each condition did not complete any regular tutor problems.

For the duration of total pre-instruction time, the interaction was not significant so the model was re-run without it. In both models (with and without the interaction term) pre-test score was significant at p=.028, achievement level was not significant at p>.17, and condition was significant at p<.03. Estimated marginal means from the main-effect model show the Grounded condition took about 4.5 minutes less for the pre-instruction than the Guided condition (18 minutes and 6 seconds vs. 22 minutes and 41 seconds, evaluated with pre-test score at .35).

To determine if the additional guided instruction caused a difference in students' learning from the tutor, I ran ANCOVAs on post-test score and delayed post-test score. The ANCOVAs included pre-test score as a covariate, with class achievement level and condition as fixed factors, and a condition by achievement interaction term. For post-test score, the interaction was not significant so the test was re-run without it. In both models, pre-test score was significant (p<.0005) and condition was not (p=.13 in the main-effects model and p=.19 in the model with the interaction term). With the interaction term, class achievement level was marginal (p=.07) and without

it, class achievement level was significant (p=.038). For delayed-test score, the interaction was not significant so the test was re-run without it. In both models, pre-test score and class achievement level were both significant (p<.01) and condition was not (p>.7). These tests were repeated on the addition items alone (post and delayed addition scores as dependent, pre addition score as covariate). On the post addition scores, the interaction term was not significant. In both models, pre score was significant (p<.0005) and class achievement level was not (p>.6). In the model with the interaction term, condition was not significant (p>.1), and in the main-effects model condition was marginal (p=.08). Estimated marginal means for the main-effects model are .397 for the Grounded condition and .317 for Guided, evaluated at a pre-test addition score of .285. On the delayed addition scores, the interaction term was not significant. In both models, pre score was significant (p<.035), as was class achievement level (p<.0005), while condition was not significant (p>.4).

These results indicate that the Guided students did not benefit from the additional transition instruction. Guided students took longer on the pre-instruction overall compared with Grounded students (a difference of about four and a half minutes, marginal significance), and, likely due to the reduced time left for the rest of the intervention, completed fewer regular tutor problems. There was no significant difference in the number of hints requested by the Guided and Grounded students during the regular tutor problems, or the amount of time taken per problem, indicating that the addition instruction did no help students solve the tutor problems. Further, there were no significant differences in learning with the tutor, on the overall test or on the addition items alone, from pre-test to post-test or from pre-test to delayed test. These results do not support hypothesis 1, and it cannot be concluded that the addition instruction the Guided students received benefitted their learning. As the two conditions had no differences in learning, they will be collapsed into one condition for further analysis, and will be referred to as Grounded.

3.3.8   Results and Analysis for Hypothesis 2

To examine the differences between the correctness condition and the collapsed grounded condition (grounded and grounded with guidance together), I re-calculated the process and outcome measures, now for two conditions. Table 3.2 shows process

	Correctness	Grounded
Pre-instruction duration (min:sec)	10:09 (0:30)	20:10 (1:04)
Regular tutor problems attempted	17.6 (.53)	9.5 (.79)
Hints per regular tutor problem	2.5 (.48)	7.8 (1.0)
Time per regular tutor problem (min:sec)	1:10 (0:06)	2:44 (0:13)

Table 3.2 Process measure means for correctness and grounded conditions: duration of number of pre-instruction (tutor introduction, addition instruction, guided transition, and repetition of the problem ½ + 1/3), number of tutor problems started, number of hints per regular tutor problem, and time spent per regular tutor problem, with standard error of the mean in parentheses.

measures by condition (time taken on the pre-instruction, number of regular tutor problems attempted, hints requested per regular tutor problem, and amount of time taken per regular tutor problem. Figures 3.13 and 3.14 show the mean scores at each test time, for the assessments overall and for the addition items, respectively.

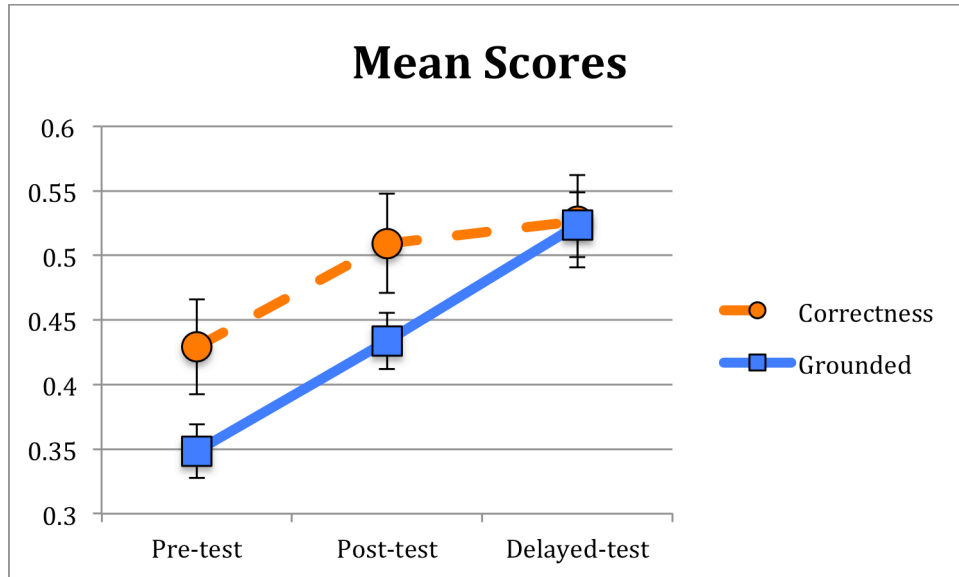**Fig. 3.13** Mean scores on each overall assessment, by condition, per test time, with bars showing standard error of the mean.
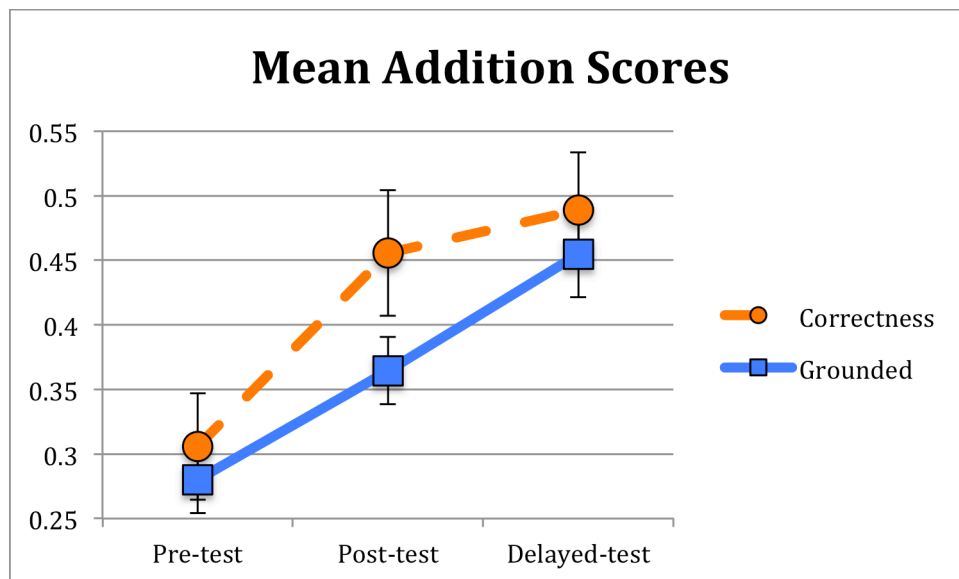

Fig. 3.14 Mean scores on the addition items, by condition, per test time, with bars showing standard error of the mean.

Process Differences. To determine if the differences in the process measures are significant, I ran ANCOVAs on the pre-instruction duration, number of regular tutor problems attempted, hints requested per regular tutor problem, and time per regular tutor problem, with condition and class achievement level as fixed factors and pre-test score as a covariate, and a condition by achievement level interaction term. For the pre-instruction duration, the interaction term was not significant so the model was re-run without it. With the main-effects model, condition was significant (p<.0005) as was pre-test score (p=.044). Class achievement level was marginal (p=.089). Estimated marginal means were 10 minutes and 59 seconds for Correctness, and 19 minutes and 47 seconds for Grounded (evaluated at a pre-test score of .377). For the number of tutor problems attempted, there was a significant effect of condition (p<.0005) and pre-test score (p=.006), with a marginal class achievement level by condition interaction (p=.050), and a marginal effect of class achievement level (p=.069). Estimated marginal means by condition and class achievement level were 17 problems attempted for Correctness (the range across all classes was 16.9 to 17.4) and, for Grounded, 7.4, 8.4, and 13.6 problems attempted for the low, middle, and high classes, respectively (evaluated at a pre-test score of .377). For the hints requested per regular tutor problem, the condition by class achievement level interaction was not significant, so the model was re-run without it. With a main-effects model, condition was significant (p=.002), pre-test score was marginal (p=.061), and class achievement level was significant (p=.012). Estimated marginal means were 3.6 hints requested per regular tutor problem for Correctness, and 7.8 for Grounded (evaluated at a pre-test score of .38). For the time taken per regular tutor problem, the condition by class achievement level interaction was not significant, so the model was re-run without it. With a main-effects model, condition was significant (p<.0005), as was pre-test score (p=.015). Class achievement level was not significant (p=.19). Estimated marginal means are 1 minute 20 seconds per problem for Correctness, and 2 minutes 40 seconds for Grounded (evaluated at a pre-test score of .38).

Differences in learning: overall scores. To determine if the different conditions led to differences in learning, I ran ANCOVAs on the pre, post and delayed scores. First, I ran an ANOVA to check for differences at pre-test, with class achievement level, condition, and pre-test form as fixed factors. Interactions were not significant in a full factorial model. A main-effects model showed that class achievement level was significant (p<.0005) as was pre-test form (p=.005). Condition was not significant (p=.111). Since test form was significant, it will be included in analyses that examine learning.

To examine immediate learning, I ran an ANCOVA on the post-test scores, with pre-test score as a covariate, and class achievement level, condition, pre-test form, and post-test form as fixed factors, with all two-way interactions for the fixed factors and a pre-test by condition interaction term. None of the interactions were significant. With a main-effects model, pre-test score was significant (p<.0005), as was class achievement level (p=.004). Condition, pre-test form, and post-test form were not significant (p>.4).

To examine retention and future learning, I ran an ANCOVA on the delayed-test scores, with post-test score as a covariate, and class achievement level, condition, delayed-test form, and post-test form as fixed factors, with a full factorial model on the fixed factors. None of the interactions were significant (another model, with all two-way interactions for the fixed factors also showed that none of the interactions were significant). In a main-effects model, post-test score and class achievement level are significant (p<.0005), as is condition (p=.036). Post-test form was not significant (p>.4), but delayed-test form was significant (p=.047). Estimated marginal means for delayed score by condition are .470 for correctness and .523 for grounded, evaluated at a post-test score of .460. The estimated marginal means for delayed score are .470 for correctness and .532 for grounded (evaluated at a post-test score of .460). Estimated marginal means for the delayed post-test form were lowest for students with a mix of questions from more than one test form (.433) and highest for students with the C form (.573).

To determine if there was a significant difference in test form distribution between the two conditions, I ran Chi-Square tests. A Chi-Square test on the distribution of forms for the fraction addition portion of the test shows no significant difference between the conditions (p>.2), as does a Chi-Square test on the distribution of forms for the other test questions (p>.2). Students were intended to have the same test form for both parts of the test (e.g., the form A for the fraction addition items was meant to always be paired with the form A for the other test items) but some students received mixed test forms, for example with form A for the fraction addition items and form B for the other test questions. Categorizing those cases as Mixed, a Fisher Exact test on the test forms for the delayed tests overall again shows no significant difference in test form distribution between the two conditions (p>.5).

To examine learning across the entire study, I ran an ANCOVA on the delayed-test scores, with pre-test score as a covariate, and class achievement level, condition, delayed-test form, and pre-test form as fixed factors, with a full factorial model for the fixed factors. None of the interactions were significant. With a model including all two-way interactions of the fixed factors and a condition by pre-test interaction term, there was a marginal effect for the condition by pre-test interaction term (p=.085) but none of the other interactions were significant. With a model that included main effects and a condition by pre-test interaction term, the interaction term and the pre-test and delayed-test forms were not significant (p>.15); class achievement level and pre-test score were significant (p<.005), as was condition (p=.013). With a model that included class achievement level, condition, and pre-test score as main effects, class achievement level and pre-test score were significant (p<.0005), as was condition (p=.039). Estimated marginal means were .469 for Correctness and .532 for Grounded (evaluated a pre-test score of .377).

Since there was no significant difference in addition learning from pre-test to delayed-test between the conditions (see below), I ran a MANOVA on the remaining parts of the test to investigate on which section grounded students were improving more. Questions were organized into groups based on pre-existing hypotheses of what the questions were assessing: pre-requisite knowledge, evaluation of fraction addition equations, and transfer items from standardized tests. With the number of questions

correct for each sub-group at delayed-test as the dependent measures and the number correct at pre-test as covariates, and condition, pre-test form, delayed-test form, and class achievement level as fixed factors, a full-factorial model shows that none of the interactions are significant at the multivariate level (Pillai's Trace). With main effects only, condition is significant at the multivariate level (p=.013), as are class achievement level (p<.0005), delayed-test form (p=.002), and number of pre-test questions correct on the evaluation and transfer items (both p<.025). Number of questions correct on the pre-requisite items (p>.2) and pre-test form (p>.7) were not significant. Tests of between-subject effects show a significant effect of condition on delayed-test transfer items (p=.002), and a marginal effect on evaluation items (p=.077). Test form for the delayed test (A, B, C, or Mixed) had a significant effect on transfer items (p<.0005) but not on the other question groups. Estimated marginal means for the evaluation questions at post-test were 1.36 for Correctness and 1.62 for Grounded (out of a maximum of 3, evaluated at a pre-test score of 1.12), and estimated marginal means for the transfer items were 1.49 for Correctness and 1.99 for Grounded (also out of a maximum of 3, evaluated at a pre-test score of 1.5).

Differences in learning: addition scores. I repeated these analyses on the addition scores alone, to determine if there was a difference in learning by condition for the content the tutors were intended to target. First, I ran an ANOVA to check for addition differences at pre-test, with class achievement level, condition, and test form as fixed factors (full-factorial model). None of the interactions were significant, so I re-ran the model with main effects only. In both models, class achievement level was significant at p<.0005, condition was not significant (p>.5) and test form was marginal (p=.083 in the full-factorial model and p=.067 in the main effects model).

To examine immediate learning, I ran an ANCOVA on the addition post-test scores, with class achievement level, condition, pre-test form and post-test form as fixed factors, and pre-test addition score as a covariate. With the full-factorial model for the fixed factors, none of the interactions were significant. In a model with all two-way interactions for the fixed factors and a condition by pre-test addition score interaction term, there was a marginal interaction between condition and class achievement level (p=.09), but none of the other interactions were significant. In a model with main effects and the condition by class achievement level interaction, there is a significant effect of class achievement level (p=.029) and pre-test addition score (p<.0005), and a significant interaction of class achievement level and condition (p=.008), without a significant effect of condition (p=.323). Estimated marginal means for the low, middle, and high classes are .306, .341, and .573 for the Correctness condition and .355, .377, and .369 for the Grounded condition (evaluated at a pre-test addition score of .289). In other words, the Grounded condition was better than Correctness for the students in the low and middle achievement classes, but worse for students in the high achievement classes. Estimated marginal means for the two conditions overall are .407 for Correctness and .367 for Grounded, (evaluated at a pre-test addition score of .289). Parameter estimates for the condition by class achievement level interaction term are -.254 for the interaction of Correctness and the lowest class level, and -.240 for the interaction of Correctness and the middle class

level. These parameter estimates, like the estimated marginal means, imply that, while there were no significant overall condition differences, the Grounded condition was better than Correctness condition for the low and middle classes, while the reverse was true for the high classes. The parameter estimate for condition is .204 for Correctness.

To examine retention and future learning, I ran an ANCOVA on the addition delayed-test scores, with class achievement level, condition, delayed-test form and post-test form as fixed factors, and post-test addition score as a covariate. With a full-factorial model, there was a marginal effect of the interaction between post-test form and delayed-test form (p=.076). Re-running the model with main effects and the interaction of the two test forms, class achievement level and post-test addition score were significant (p<.0005) with no significant effect for condition (p>.2), the test forms (p>.1), or the interaction between the test forms (p>.2). Condition remains not significant with main effects only (p>.2) and with a model that does not include the test forms (p>.1). Estimated marginal means for the model with main effects only, including test forms, are .417 for Correctness and .465 for Grounded, evaluated at a post-test target score of .397.

To examine addition learning over the entire study, I ran an ANCOVA on the addition delayed-test scores, with class achievement level, condition, delayed-test form and post-test form as fixed factors, and pre-test addition score as a covariate. With a full-factorial model, none of the interactions were significant (the achievement level by condition interaction was marginal, p=.078). Re-running the model with main effects and a class achievement level by condition interaction term, the interaction term was not significant (p>.2). With main effects only, pre-test addition score and class achievement level are significant (p<.0005), with a marginal effect for the addition pre-test form (p=.078) and no significant effect for the addition delayed-test form (p>.9), and no significant effect of condition (p>.6).

3.3.9   Results and Analysis for Hypothesis 3

To determine if students in the grounded conditions understood the grounded feedback, I analyzed scores on assessment items that included fraction bars, and students' overall problem evaluations while working with the tutor.

Evaluation items. Three item on each test assessed students' ability to evaluate a proposed fraction addition equation. The problems were presented as fictitious student work, and asked if the work was correct. One item showed the addends, converted fractions, and sum with fraction bars and fraction symbols (one showed a procedural hint on converting instead of showing the fraction bars, and one showed only the problem statement and fictitious work). For the problems with fraction bars, the layout is similar to the tutor interface for the grounded conditions (Fig. 3.15). The three test forms crossed three strategies for each problem type: correct (fractions correctly converted to a common denominator and added correctly), incorrect conversion (a common denominator is used, but the numerators from the original
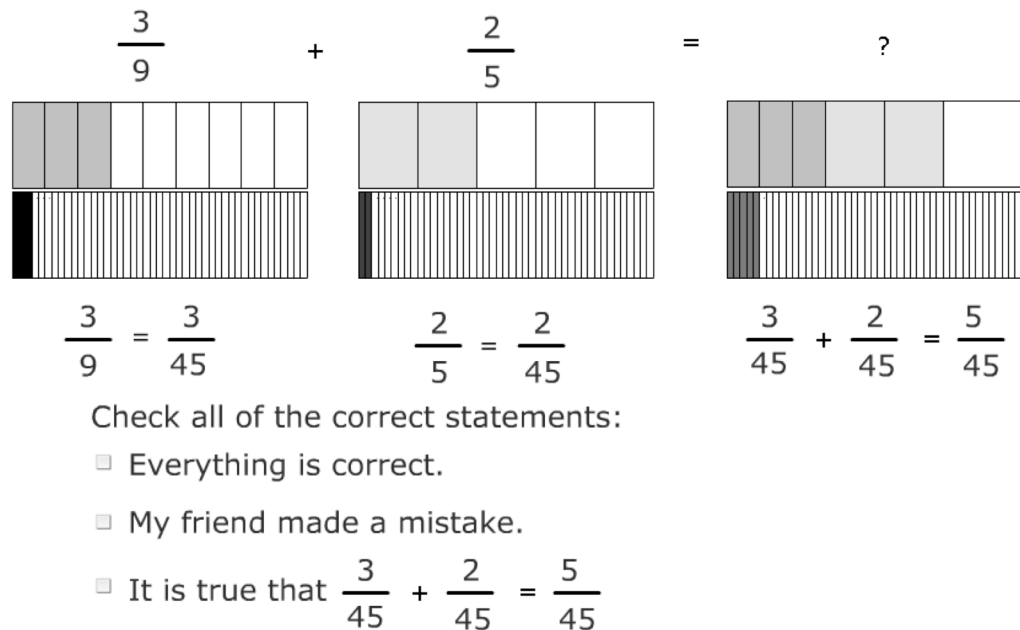
$$\frac{3}{9} \quad + \quad \frac{2}{5} \quad = \quad ?$$

$$\frac{3}{9} = \frac{3}{45} \qquad \frac{2}{5} = \frac{2}{45} \qquad \frac{3}{45} + \frac{2}{45} = \frac{5}{45}$$

Check all of the correct statements:

☐ Everything is correct.

☐ My friend made a mistake.

☐ It is true that $\frac{3}{45} + \frac{2}{45} = \frac{5}{45}$

Fig. 3.15 Evaluation item with magnitudes represented as fraction bars.

addends are retained; these quantities are then added correctly), and add-both (no conversion, and the sum is obtained by adding the numerators and denominators independently). Students were asked to evaluate if the fictitious work is completely correct or if there was a mistake in it, and were also asked if the addition of the 'converted' fractions was correct.

On the evaluation items with fraction bars, the proposed sum fraction only lines up with the combined magnitude of the two addends when the equation is correct. With the 'incorrect conversion' strategy, the proposed sum is less than the first addend, and with the 'add-both' strategy, the proposed sum is about half of the true sum. This section considers only the first task, determining if everything is correct or if the friend made a mistake. For students who understand the task directions and who recognize that a sum is equal to the combined magnitudes of its addends, the task should be trivial. However, student scores on this problem were low at pretest (.64 for Correctness, .48 for Grounded). Figure 3.16 shows the mean score on this item per condition, for each test time. At pre-test, the Grounded students performed no better than chance, suggesting that the rectangles in the tutor would not be sufficient for students to evaluate if their work was correct or not. These results indicate that students started the study without sufficient prior knowledge to take full advantage of the fraction bars in the tutor. Further, the presence of the rectangles on the evaluation items may not have made those items easier than the no-scaffold evaluation items. While the mean score on the pre-test rectangle evaluation item was .54, and the mean score on the no-scaffold evaluation item was .48 (across conditions), the difference in performance across the two items was not significant (p=.332 for McNemar's test, exact sig., 2-sided, on the cross tabulation of performance on each item at pretest). One may hypothesize that students in the

grounded condition may have learned to interpret the rectangles over the course of using the tutor. However, for students in the grounded condition, while performance on this task improved from pre-test to post-test, the difference was not significant (mean score at pretest was .48, mean score at post-test was .59, p=.211 for McNemar's test). Yet, over the course of the whole study (pre-test to delayed-test), students in the grounded condition improved on this item while students in the correctness condition did not (p=.001 for McNemar's test for the grounded condition, and p>.8 for the control. Mean scores at pre-test and delayed-test were .48 and .74 for grounded and .65 and .60 for correctness, respectively).
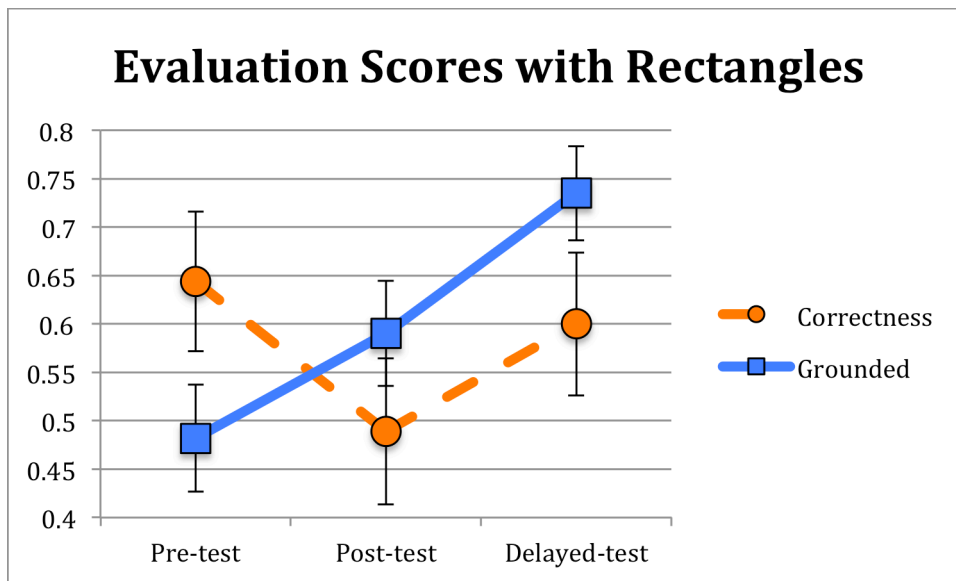

Fig. 3.16 Scores on students' evaluations of a fictitious student's work (if everything was correct or if the student made a mistake). Students saw one item of this type per assessment. Bars show standard error of the mean.

Process Measures. Students pressed the 'done' button in the tutor interface to get problem-level feedback, moving on to the next problem if the current one was solved correctly, or getting a text prompt if something needed to be fixed. The text prompts indicated that the problem was not solved correctly and encouraged the student to ask for a hint if they weren't sure what to do. If students correctly interpreted the grounded feedback, they would not press the 'done' button when their proposed sum did not line up with the correct sum. However, even in cases where the student's proposed sum differed from the correct sum by more than .1, students in the Grounded condition pressed the 'done' button .99 times per student-problem (on average). In contrast, correctness students proposing a sum that differed from the correct sum by more than .1 pressed the 'done' button .01 times per student-problem (on average). Overall, students in the Grounded condition pressed the 'done' button incorrectly 2.39 times per student-problem, contrasting with the .23 times per

student-problem in the Correctness condition (note these values include all students who used the tutors during the study, even if they did not take the delayed-test).

3.3.10 Results and Analysis for Hypothesis 4

To determine if initial problem sequence affected students' learning, I ran an ANCOVA on post-test scores with sequence, condition, pre-test form, post-test form, and class achievement level as fixed factors and pre-test score as a covariate. In a full-factorial model, none of the interactions were significant. Re-running the model main effects only, pre-test is significant ($p<.0005$), as is class achievement level ($p=.005$) with no other significant main effects.

To determine if initial problem sequence affected students' future learning and retention, I ran an ANCOVA on delayed-test scores with sequence, condition, delayed-test form, post-test form, and class achievement level as fixed factors and post-test score as a covariate. In a full-factorial model, there was a significant interaction between post-test form and sequence ($p=.041$). Re-running the model with main effects and the post-test form by sequence interaction, the interaction is no longer significant ($p>.2$). There is a significant effect of post-test score ($p<.0005$), as was class achievement level ($p<.0005$) and condition ($p=.019$), with a marginal effect for sequence ($p=.073$). There was a significant effect of delayed-test form ($p=.036$) but not of post-test form ($p>.1$). Estimated marginal means were .471 for Correctness and .540 for Grounded, evaluated at a post-test score of .460. Estimated marginal means by sequence are .448 for Same-Unrelated-Multiple, .528 for Same-Multiple-Unrelated, and .545 for Random, also evaluated at a post-test score of .460.

To determine if sequence affected learning across the duration of the study, I ran an ANCOVA on delayed-test scores with sequence, condition, delayed-test form, pre-test form, and class achievement level as fixed factors and pre-test score as a covariate. In a full-factorial model, there is a marginal three-way interaction between sequence, class achievement level, and delayed test form ($p=.05$). With a main effects model and the three-way interaction, the interaction is no longer significant ($p>.2$). Pre-test score and class achievement level are significant (both $p<.0005$), condition is marginal ($p=.062$), and sequence ($p>.8$) and the test forms are not significant ($p=.19$ for delayed-test form, and $p>.2$ for pre-test form). With main effects only, pre-test score and class achievement level are significant ($p<.0005$), as is condition ($p=.006$). Sequence ($p>.1$) and pre-test form ($p>.4$) are not significant. There is a marginal effect for delayed-test form ($p=.083$).

3.3.11 Discussion and Limitations

The results and analyses indicate that, for immediate (pre-to-post) learning of fraction addition, the grounded condition was more beneficial for the low and middle classes, while the correctness condition was more beneficial for the high classes. Overall, both conditions had similar improvement on the fraction addition items from pre-test to post-test and from pre-test to delayed-test. On the assessments as whole,

there was similar improvement for both conditions from pre-test to post-test, but greater improvement for the grounded condition from post-test to delayed-test and from pre-test to delayed-test. This greater improvement for the grounded condition on the tests overall between pre-test and delayed-test appears to be driven by greater improvement on the transfer items, and, to a lesser extent, greater improvement on the evaluation items. Students benefitted more from grounded feedback even though the grounded feedback tutor was more difficult than the correctness feedback tutor (students in the grounded condition solved fewer problems, took longer on each problem, and requested more hints per problem). These result indicate that grounded feedback is a desirable difficulty (Bjork & Bjork, 2009).

However, these results should be interpreted cautiously. Due to an error in student randomization, more students from the higher-level classes were assigned to the correctness condition. While there are no significant differences by condition at pre-test when class level is in the model, an ANOVA on pre-test scores with condition and pre-test form alone (without class level) does show a significant effect of condition (p=.012, with higher estimated marginal means for the correctness condition). Therefore, while the grounded condition shows greater improvement from pre-test to delayed-test, the grounded condition does not show better performance at delayed-test (that is, an ANOVA on delayed-test scores with condition, class achievement level, and delayed-test form as main effects –but without pre-test score in the model– shows no significant effect for condition, p>.14; removing class achievement level from the model, condition remains not significant, p>.9). Students in the grounded condition were behind at pre-test but caught up by the delayed-test. One interpretation of these results is that grounded feedback led to better learning. However, since performance at delayed-test was not greater for the grounded condition, and since there was not greater improvement between pre-test and post-test, one cannot rule out the hypothesis that the two weeks of classroom instruction between the post-test and delayed-test simply brought all students up to the same level.

Further, while there was not a significant difference in the *number* of students who started but did not complete the study between the two conditions, there may be a difference in the amount of learning demonstrated by those students. Four students in the correctness condition did the pre-test, worked with their tutors, and did the post-test but did not do the delayed-test. Of those four students, three demonstrated improvement in fraction addition between the pre-test and post-test. Six students in the grounded condition also did all parts of the study except for the delayed-test. Of those six, only one demonstrated improvement on the fraction addition items between the pre-test and post-test. These students were excluded from the analyses above because they did not complete all parts of the study. Including these students in an ANCOVA with the same model as used in the analyses above (pre-test addition scores as a covariate, and condition, class achievement level, and the test forms for each test time as fixed factors, and a condition by class level interaction term) shows a marginal effect of condition (p=.075), with estimated marginal means of .283 for correctness and .214 for grounded (evaluated at a pre-test addition score of .283). Greater immediate learning for the correctness condition is consistent with cognitive

load theory, especially because, in the correctness condition, students' attention is not split between interpreting the fraction bars and practicing the procedure.

These results also suggest that the initial question sequence may have had some effect on students' learning, with a marginal difference for learning between the post-test and delayed test, but no significant difference for immediate learning (pre-test to post-test) or learning over the whole study (pre-test to delayed-test). Question sequences may affect learning more strongly when they differ across the entire intervention, not just in the first few problems.

These results do not support the hypothesis that students working with the grounded feedback tutor benefited from the brief instruction relating the fraction bars to the concrete representation of bugs and rulers. This instruction may have been too brief to affect students' interpretations of the fraction bars. Further, pre-test assessments and process measures indicate that students in the grounded feedback condition were not using the fraction bars effectively in evaluating if a fraction addition equation was correct or not. At pretest, the inclusion of the fraction bars did not make the evaluation task easier than the unscaffolded, numbers-only task, and performance in the grounded condition was around chance. While working with the tutors, students in the grounded condition frequently indicated that their work was correct, even when their proposed sum did not line up with the rectangle that represented the combined magnitudes of the two addends. Although students in the grounded condition learned, and learned more over the course of the study than the correctness students, they seemed to not fully understand the grounded feedback.

3.4   Conclusion

This chapter presented the grounded and correctness feedback tutors, and a controlled *in vivo* study comparing them. In terms of overall learning across the duration of the study, grounded students improved more than correctness, with no differences in learning on the target fraction addition content. This suggests that grounded feedback helps students learn fraction addition as well as a high-bar, symbols-only correctness tutor, even though students in the grounded condition had to transfer their learning across representations. These outcome measures indicate that while students in the grounded condition struggled more while using the tutor, those difficulties were desirable. However, students in this this study did not interpret the fraction bars with the same apparent ease as students in the pilot, demonstrated with near-chance performance on an assessment item at pretest and students' incorrect 'done' presses during the intervention.

This chapter provides evidence that middle school students can benefit from grounded feedback for learning fraction addition. Also, while this study was not designed to examine the features of grounded feedback individually, it does provide evidence for the benefits of using an intrinsic representation as feedback during learning. This feature was not the only difference between the conditions – the correctness condition had immediate correctness feedback and the grounded condition did not. However, it is unlikely that the learning benefits for the grounded

condition came from the lack of correctness feedback rather than the presence of the fraction bar feedback.

4   Evaluating How Students Relate Magnitude to Addition with Difficulty Factors Assessments

Summary. What types of scaffolds support sense making in mathematics? Prior work has shown that grounded representations such as diagrams can support sense making and enhance student performance relative to analogous tasks presented with more abstract, symbolic representations. For grounded representations to support students' learning of symbolic representations, students' sense making must be maintained when both grounded and symbolic representations are presented together. This study investigates why students sometimes fail to coordinate these representations, in particular, why performance is high with fraction diagrams alone, but decreases when fraction symbols are included. Results indicate that symbols trigger incorrect transfer from whole-number procedures, and that students lack the qualitative reasoning that the diagrams are intended to tap. Specifically, students do not find it obvious that the sum of two positive symbolic fractions is larger than its two addends. Qualitative inference rules such as this one appear important in mediating the sense making process in the context of tempting misconceptions even when otherwise-supportive grounded representations are available.

4.1   Motivation

Many researchers strive to identify ways to support deep understanding, as it is thought to promote robust and adaptable learning. One strategy has been to use multiple representations, particularly ones that connect to students' prior knowledge and aid sense making. One way to reinforce the conceptual foundation for procedures

is to use visual representations, such as strip diagrams. These diagrams are not intended to help student execute procedures, but instead support them in thinking about the problems qualitatively (e.g., which amounts are bigger? Which operation is appropriate?). Visual representations are thought to help students apply their conceptual reasoning (Beckmann, 2004), and are recommended by an Institute of Education Sciences Practice Guide (Woodward et al., 2012). However, there is little data on what representations will make sense to the students. Further, diagrams may not be intuitive for novices, and their presence can decrease problem-solving performance for students who have difficulty interpreting them (Booth & Koedinger, 2011). Booth and Koedinger (2011) hypothesized that several factors could contribute to students' misinterpretation of diagrams or their difficulty mapping between diagrams and problem statements, including a lack of domain knowledge and still-developing formal reasoning.

Results from the previous chapter suggest that students have difficulty using fraction bars to evaluate if a fraction addition equation is correct. This finding revealed that the fraction bar representations of addition were not as meaningful to all students as the think-aloud results from Chapter 2 suggested. Thus, this chapter investigates more deeply the cognitive mechanisms required for processing these representations and, in particular, attempts to identify the sticking points where student processing deviates from expectation.

4.2   DFA Study 1: Evaluating Equations

This difficulty factors assessment (cf., Koedinger, Alibali, & Nathan, 2008) examines how students understand fraction bars in the context of the fractions they represent; if this process changes depending on the topic (addition vs. equivalence); and how each processing step affects performance.

4.2.1   Cognitive Task Analysis and Test Items

A theoretical cognitive task analysis identified three likely skills needed to understand the fraction bar representations for fraction addition: 1) equal areas represent equal amounts; 2) the rectangular bars represent the symbolic fractions written above or below them; 3) if two shaded areas are equal, the fractions they represent are equal. The first skill addresses students' interpretation of the fraction bars on their own, while the second and third skill addresses students' coordination of the fraction bar and fraction symbol representations. I developed matched test items intended to isolate those skills (Fig. 4.1 – 4.4). Fraction addition items presented a fully solved problem and students indicated whether it was solved correctly (true or false). Fraction equivalence items presented two fractions and students indicated if the first fraction was bigger than, equivalent to, or smaller than the second fraction. The four question presentations are intended to isolate the skills needed to make sense of the grounded tutor interface in Chapter 3. The pictures format (Fig. 4.1) assesses if students know that the shaded rectangles use area to represent quantity, such that

two rectangles with equal-sized shaded areas represent equal quantities. Pictures-and-numbers items (Fig. 4.2) include fraction symbols with the fraction bars, to test if students can understand the fraction bars as representations of fractions. Half-pictures-and-numbers items (Fig. 4.3) also include both fraction bars and fraction symbols, but only present the fraction bars as the hint at the top of the problem. This determines if students can find the relationship between the two fraction bars, map that relationship to the symbolic fractions represented, and then select the relationship that the symbolic fractions have to each other. Numbers-only (Fig. 4.4) provides a baseline for how well students can evaluate the equivalence and addition problems without fraction bars. Another pair of questions gives a baseline for translating a single fraction bar to a fraction symbol (e.g., when shown a rectangle divided in 6 parts with 4 of them shaded, the student should write 4/6; Figs. 4.5, 4.6)

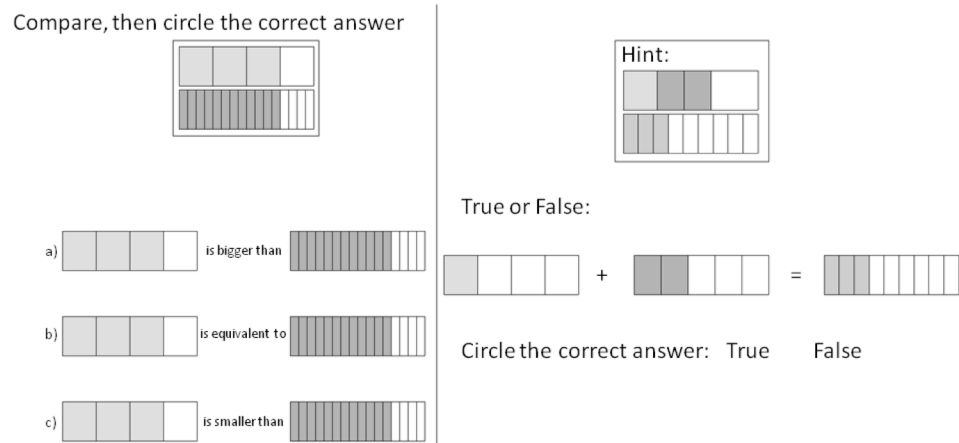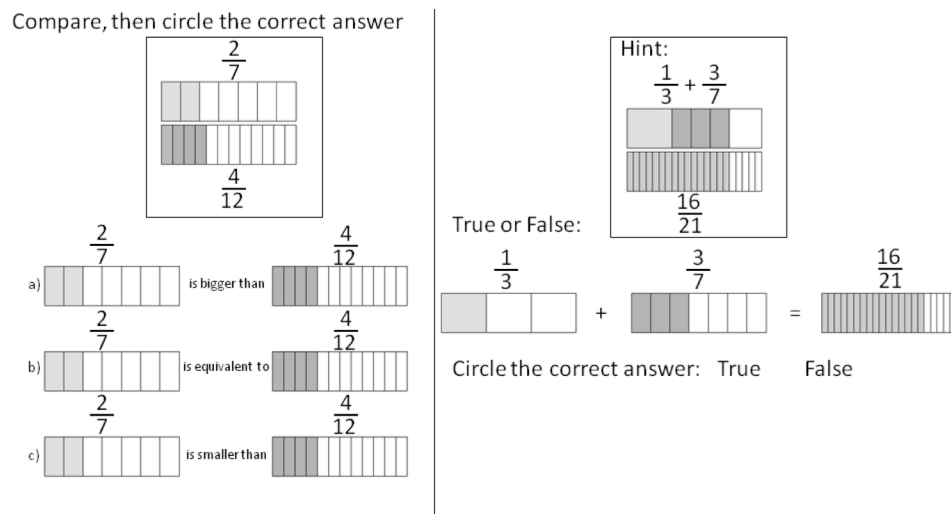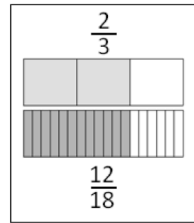**Fig. 4.1** Pictures-only items.

Fig. 4.2 Pictures and Numbers items.

Compare, then circle the correct answer

$$\frac{2}{3}$$

$$\frac{12}{18}$$

Hint:

$$\frac{3}{4} + \frac{1}{7}$$

$$\frac{4}{11}$$

a) $\frac{2}{3}$ is bigger than $\frac{12}{18}$

b) $\frac{2}{3}$ is equivalent to $\frac{12}{18}$

c) $\frac{2}{3}$ is smaller than $\frac{12}{18}$

True or False:

$$\frac{3}{4} + \frac{1}{7} = \frac{4}{11}$$
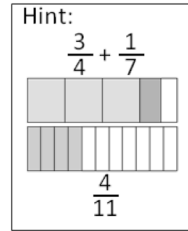
Circle the correct answer: True   False

Fig. 4.3 Half Pictures and Numbers items.

Compare, then circle the correct answer

a) $\frac{1}{3}$ is bigger than $\frac{8}{19}$

b) $\frac{1}{3}$ is equivalent to $\frac{8}{19}$
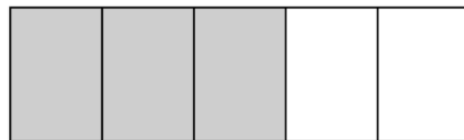
c) $\frac{1}{3}$ is smaller than $\frac{8}{19}$

True or False:

$$\frac{2}{11} + \frac{1}{2} = \frac{15}{22}$$

Circle the correct answer:
True   False

Fig. 4.4 Numbers-only items.

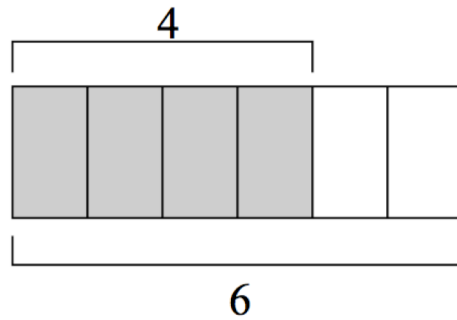What fraction does this represent? _____

Fig. 4.5 Single fraction bar.

4.2.2 Participants

155 fifth grade students from a local public school participated in this study during their normal school day (the same school as the study in Chapter 3, during a different school year). The school tracked these classes, with 57 students in the highest track, 61 in the middle track and 37 in the lowest track.

What fraction does this represent? _____

Fig. 4.6 Single fraction bar with numbers.

4.2.3 Materials and Method

Paper test forms included 8 equivalence items and 8 addition items (one correctly solved and one incorrectly solved for each scaffold type). All addends in these items had unlike denominators. The sums in the incorrect addition items followed the common misconception of adding both numerators and both denominators. Tests also included two single fraction bar items, one with numbers for how many pieces were shaded and how many total. Item presentations were counterbalanced with the specific numbers in the problems to avoid confounding. These items were embedded in a 30-item assessment, with item order was determined randomly and half of the tests printed with the question order reversed. Students were given 20 minutes for the full 30-item assessment. Questions were scored 1 if correct and 0 otherwise.

4.2.4 Results: Scaffold Type Affects Performance

Scores on the single-fraction-bar items were near perfect (94% correct). For the addition items, some questions were misprinted on some of the test forms. Therefore, the following analysis only includes 122 participants for the pictures-only addition questions and 140 students for the pictures & numbers addition questions. Figure 4.7 shows the mean scores for the equivalence and addition items by scaffold type. Mean scores on the fraction equivalence items were high, with 81-83% correct for all scaffold types with pictures, and 50% for the numbers-only presentation. Equivalence items offered three options (bigger, equivalent, or smaller) so even the numbers-only score is well above 1/3 chance. Mean scores on the fraction addition items were lower (21% to 79%). These scores steadily decreased as the saliency of the numbers increased. Lower-than-chance results indicate that instead of guessing randomly on the more difficult scaffolds, students answered based on a systematic misconception. Blank

answers were scored as 0 and they could reduce performance below the 50% chance rate. However, students were no more likely to skip the numbers-only
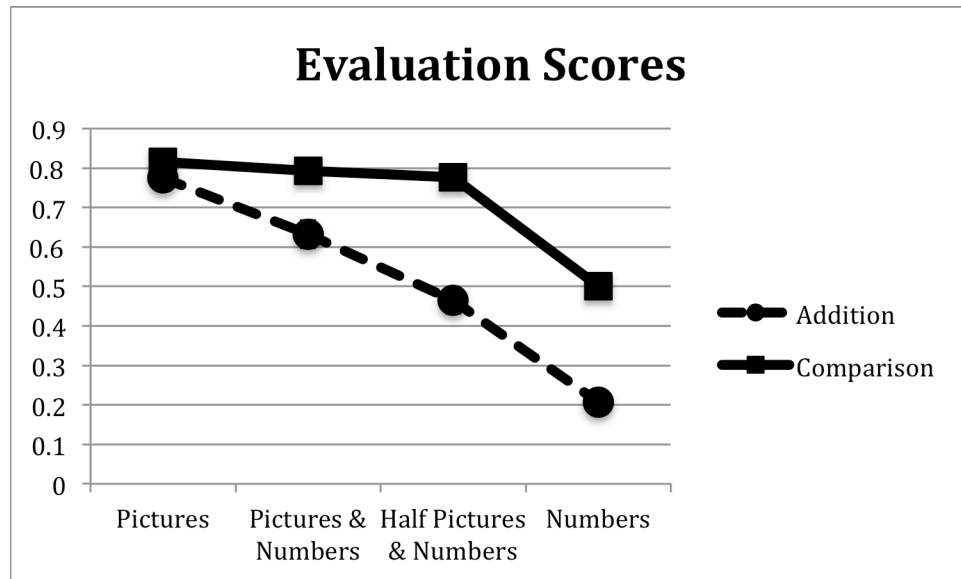
Evaluation Scores

Fig. 4.7 Mean scores on the comparison and addition evaluation items, by scaffold type. Error bars showing the standard error of the mean are mostly covered by the point markers.

addition items than the other addition items that included numbers (numbers-only addition was skipped 13 times, while half-pictures-and-numbers and pictures-and-numbers were skipped 14 times each).

There is a strong interaction effect between question type and scaffold type. I ran an ANOVA on the item scores: 3 (class tracking level: high, middle, low) x 4 (scaffold type: pictures, pictures and numbers, half pictures and numbers, numbers only) x 2 (item: equivalence or addition) with repeated measures for the scaffold type and item. With the Huynh-Feldt correction (since sphericity could not be assumed), results showed significant within-subjects effects for scaffold type and item, and a significant scaffold by item interaction (all p<.0005). Results also showed significant between-subjects effects for class tracking level, with parameter estimates indicating that higher-tracked students got higher scores.

The patterns in figure 6 suggest that all scaffold types with pictures have a similar effect for equivalence, but each scaffold type has a different effect for addition. To verify these hypotheses statistically, I ran separate ANOVAs on each tracking level for equivalence and addition scores, with scaffold type as a fixed factor and student as a random factor. For each of those analyses on the equivalence scores, scaffold was significant (p<.0005) and post-hoc Tukey tests showed that the numbers-only scaffold was significantly different from the other three (p<.0005). For each of those analysis on the addition scores, scaffold was again significant (p<.0005). Tukey tests for the middle track show significant

differences among all scaffold types (p<.01). The lowest track did not have significant differences between half-pictures-and-numbers and numbers-only, likely a floor effect. The highest track did not have significant differences between pictures and pictures-and-numbers, likely a ceiling effect.

Figure 4.7 also suggests that addition with the pictures-only scaffold is no more difficult than equivalence with the pictures-only scaffold. To test this, I ran an ANOVA on the item scores for the pictures-only scaffold: 3 (class tracking level: high, middle, low) x 2 (item: equivalence or addition) with repeated measures for item. Results showed no significant difference for scores on the two question types (p = .2 with the Huynh-Feldt correction). Subsequent ANOVAs on each of the other scaffold types showed significant differences for scores on the two question types (all p<.0005 with the Huynh-Feldt correction).

One may hypothesize that when pictures are present, students would be more accurate when there is a large disparity in the area of the quantities being compared. To test this hypothesis, I calculated a disparity measure for each question where the two fractions were not equivalent or the two addends did not equal the sum. For the equivalence items, the disparity is the absolute value of the first fraction minus the second fraction. For the addition items, the disparity is the true sum of the addends minus the sum in the question. I ran separate ANOVAs for each question type, with scaffold type and disparity as fixed factors and student ID as a random factor. For both addition and equivalence, between-subject main effects were significant for scaffold type and student ID (p<.0005) but not for disparity (p=.141 for addition, p=.888 for equivalence), and there was no scaffold*disparity interaction (p=.257 for addition, p=.136 for equivalence). This indicates that disparity did not affect scores, and the effect of disparity did not change with scaffold type. Additionally, the equivalence questions all had smaller disparities than the addition questions (means: .06 for equivalence, .39 for addition), yet the equivalence questions were as easy or easier, further evidence that disparity did not affect scores.

4.2.5   Discussion: Fraction Bar Skills are Context-Based

Section 4.2.1 hypothesized three skills that were necessary for making use of the fraction bars for fraction addition: 1) interpreting the colored areas as amounts that can be compared, such that equal areas represent equal amounts; 2) relating the fraction bars to the fraction symbols, such that the fraction bars are interpreted as representing the magnitudes of their corresponding fraction symbols; 3) coordinating the fraction bars and fraction symbols, such that if a relationship is present between two fraction bars, that same relationship is applied to the fraction symbols (e.g., if two fraction bars show equal areas colored in, the fractions that they represent are also equal). The results from the difficulty factor assessment indicate that some of these skills are context-dependent. Students were at ceiling for writing the symbolic fraction represented by a single fraction bar, whether or not numeric symbols were included, indicating ease with interpreting a fraction bar on its own. In both the comparison and addition

contexts, students had high performance on the pictures-only task, demonstrating ease with the first skill (equal areas represent equal amounts). However, for the second and third skills, performance differed by context: the presence of fraction symbols with the fraction bars reduced performance in the fraction addition context and had no significant effect on performance in the comparison context. This finding indicates that the second skill, seeing the fraction bars as representations of fraction symbols, poses a difficulty in fraction addition but not in fraction comparison. The half-pictures-and-numbers task, which required students to coordinate between the two representations, again posed an additional difficulty with fraction addition and had no significant effect with fraction comparison. If students' performance with the numbers-only comparison task was also high, one could conclude that students were simply solving the comparison problems based on the fraction symbols. However, performance on both tasks was significantly lower with the numbers-only format. Still, since the comparison task gave three multiple-choice options and the addition task gave two, performance was above chance for the numbers-only comparison task and below chance for the numbers-only addition task. These results show that even though fraction bars in a fraction addition context make it easier for students to evaluate if an equation is correct or not, students still have substantial difficulty coordinating the fraction bar and fraction symbol representations.

I hypothesize that the interference of the incorrect add-both-numerators-and-denominators strategy overrides the area-as-quantity reasoning that students demonstrate when the numbers are not shown. A cognitive-load hypothesis may predict that fraction symbols are distracting because they visually clutter the problem. In that case, scores with half pictures and numbers should be higher than pictures and numbers, since there is less information and less visual clutter. Yet, scores decrease, indicating that performance is not correlated with cognitive load. Byrnes and Wasik (1991) discuss a theory that conceptual knowledge will prevent students from making certain procedural errors. In this theory, a "self-critic" (my name), evaluates procedural outcomes for conceptual errors. For example, if a student adds 3/4 and 1/7 and gets 4/11, their "self-critic" may reason that 4/11 cannot be right because it is less than half while 3/4 is greater. With the picture scaffolds, these steps are easier – instead of numeric mental operations, students can compare the fraction bars. Scores on the equivalence and the pictures-only addition items demonstrate students' skill in comparing fraction bars, yet they still seem to not use their "critic" on the fraction addition items with numbers.

Interestingly, Byrnes and Wasik argue against the self-critic theory, claiming that conceptual and procedural knowledge are not commonly activated simultaneously in problem solving. Further, conceptual knowledge may precede procedural skill, so in some stages of learning conceptual knowledge would not be correlated with procedural performance. Instead, procedural skills improve through proper discrimination and generalization. To test these theories, they compared three instructional techniques for LCD fraction addition. One was

procedural, and stressed that "you can't add fractions the way you add ordinary numbers." The other techniques added conceptually based instruction (one with fraction bars) to that procedural instruction. Results showed that the conceptual methods did not improve learning above the purely procedural one. These findings suggest that aiding discrimination will improve procedural skill, and that skill is not enhanced further with brief conceptual instruction. However, while the conceptual method included a *demonstration* of using fraction bars and coordinating between the two representations, students did not actually *practice* this skill themselves. This difficulty factor assessment demonstrates that coordinating the fraction bars and fraction symbols is not trivial for students, and suggests that this coordination is a pre-requisite skill for the proper functioning of the self-critic, or, in other words, for the activation of conceptual knowledge in a procedural context. That is, it is not sufficient for students to have the conceptual knowledge that two positive addends result in a sum that is greater than each: students demonstrate proficiency with this concept in the pictures-only addition task. Students must also have the (metacognitive) procedural knowledge to relate that conceptual knowledge to a symbolic context. Therefore, while students will likely not benefit from further conceptual instruction on the fraction bars alone, the self-critic likely would benefit from support in coordinating the fraction bar and fraction symbol representations.

As students develop self-critic procedures, there are other potential roadblocks besides fraction misconceptions, in particular misconceptions related to the meaning of the equals sign. McNeil et al. (2006) found that 6th-8th grade students looking at a problem such as 3 + 4 = 7 were more likely to interpret the equals sign to mean "write answer here" than "both sides are equivalent." Perhaps this misinterpretation of the equals sign interferes with the application of self-critic procedures on fraction addition items. Even when the pictures show the sum to be smaller than one of the addends, the student may not realize that the two sides of the equal sign are supposed to be equivalent. A self-critic that interprets the equal sign as "write output of procedure here" may simply verify that the add-both-numerators-and-denominators strategy was executed well. In other words, the presence of numbers may not only prompt over-generalization of whole-number addition, but also interfere with students' interpretation of the equals sign and thus throw off the self-critic.

4.2.6 Conclusions

These data imply that the usefulness of the fraction bar scaffold is dependent on the topic for which it is employed. When naming fractions represented by individual fraction bars and solving equivalence problems with fraction bars, students were equally proficient whether the numeric symbols were present or not. However, for fraction addition, the presence of fraction symbols interfered with the use of the fraction bars. The pictures-only addition problems may invite reasoning based on conceptual understanding (the sum of two areas cannot be

smaller than either addend), while the presence of fraction symbols may invite procedural problem solving that is initially divorced from the underlying concepts.

This DFA study suggests that some of students' difficulty with dynamic fraction bars in the tutoring system was due to the specific addition context. More broadly, it suggests caution in the design and use of conceptual scaffolds for math problems. Students may demonstrate proficiency with a scaffold in one domain without being able to transfer those skills, even to a closely related domain. Procedural misconceptions may override the conceptual reasoning these scaffolds attempt to induce. Perhaps students need instruction to support their "self-critics" in checking procedural outcomes against conceptual knowledge. Or, perhaps students require certain domain-specific knowledge before their "self-critics" are triggered.

4.3   DFA Study 2: Replication and Extension

Results from DFA Study 1 indicate that the presence of symbols seems to detract from students' use of the diagrams in the addition problems. Students' below-chance performance (21%) with numbers-only indicates that adding the numerators and denominators is a tempting foil, as it draws on students' incorrect transfer from whole- number addition. However, the prior study did not have sufficient error-type data to confirm this suspicion. Students' performance on the comparison tasks indicates that they can extract information equally from all three fraction-bar scaffold types. What prevents them from using this information with addition? I hypothesize a sense making process that demands recognition of two basic properties of positive-number addition for effective use of the fraction bars: 1) the magnitude of the sum equals the combined magnitudes of the addends; and 2) the sum is larger than each of the addends. With this knowledge, the incorrect symbolic addition equations should be easy to reject, since all propose a sum that is smaller than one of the addends. This study examines whether students know the second, presumably more difficult, property.

The prior study also left other open questions. First, the 'true or false' options did not give any insight into students' reasoning. Second, the comparison items with non-equivalent fractions did not use foils based on possible misconceptions. Instead, they used fractions with similar magnitudes. Perhaps students have systematic misconceptions about equivalence, as they do with addition, but those misconceptions were simply not elicited. This study addresses three questions about sense making support for fractions: 1) Is it obvious to students that the sum of two positive symbolic addends is larger than each addend individually? 2) When students do not recognize the correct sum of a fraction addition equation, is it due to incorrect whole-number transfer? 3) Are students tempted by systematic foils for fraction equivalence?

4.3.1 Participants and Method

This study was conducted with the same fifth- grade public-school students as the prior study. The prior study took place in the fall and the present study took place in the spring. Thus, students had about 5-6 months more classroom instruction in the present study than the previous one, explaining their higher scores on comparable tasks. 160 fifth-graders were given 20 minutes to complete the 34-item test forms, administered by their classroom teacher during the normal school day. The school tracked students into three achievement levels, which we refer to as High, Middle, and Low. To control for ordering effects, question order was determined randomly and half of the test forms were printed in reverse order. Within each class, students were randomly assigned to one of the four test forms, printed in either forward or reverse order. Two items were inadvertently left off the test forms of 19 students, and I account for this discrepancy in the analysis.

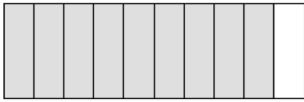## 4.3.2 Replication of DFA Study 1 with Equivalence Foils

The addition and comparison items used the same types of scaffolds as the prior study. However, this time addition items offered three responses: the sum could be too small, correct, or too big. Comparison items with non-equivalent fractions aimed to assess the extent of three potential misconceptions: fractions with the same numerator are equivalent, regardless of denominator (e.g., 3/4 and 3/16); squaring the numerator and denominator maintains equivalence (e.g., 2/5 and 4/25); and adding the same number to the numerator and denominator maintains equivalence (e.g., 11/12 and 14/15).

I refer to the foil types that target these misconceptions as *same numerator*, *squaring*, and *one-less*, respectively (one- less refers to the addition misconception since each numerator is one less than its denominator). Figure 4.8 gives an example of the half-pictures-and-numbers scaffold with the one-less foil and correct addition. This study used a between-subject design for scaffold (with each test form using only one of the four scaffold types) and a within-subject design for task (each student did comparison and addition items). Tests included 6 addition items and 12 comparison items.

19 of the 160 tests inadvertently had 11 comparison items instead of 12, so I used percent correct instead of raw scores in all analyses. An ANOVA with task (comparison and addition) as a repeated measure, and with scaffold type, tracking level, and question order (forward vs. reversed) as fixed factors showed that question order was not significant and had no significant interactions, so I re-ran the analysis without it. I found a significant effect of task (p<.01) but no significant task by scaffold interaction. For each task (comparison and addition) I ran an ANOVA on percent correct (dependent) with scaffold type and tracking level as fixed factors. For comparison items, there was a significant effect of scaffold and class level (both p<.01), with a scaffold by class level interaction (p=.013). Post-hoc Tukey tests showed significant differences between numbers-
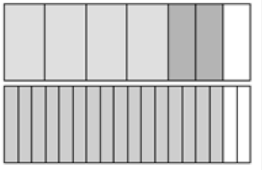
Fig. 4.8 Half Pictures and Numbers items, with one-less equivalence foil.

only and all other scaffold types (all p<.001) but no other significant differences. For addition items, scaffold and class were again significant, with a marginal interaction (p=.058). Post-hoc Tukey tests showed significant differences between numbers-only and all other scaffold types (all p<.015); pictures-only and pictures-and-numbers (p<.01); and a marginal difference between pictures-only and half-pictures-and-numbers (p=.087). Since those tests revealed no differences between the two scaffold types with both representations and did reveal differences between them and the scaffold types with one representation, I collapse those two scaffold types for further analysis. An ANOVA with the three scaffold groups and class level as fixed factors showed significant main effects (p<.01) and a significant interaction (p=.031). Post-hoc tests show significant differences between all three scaffold groups (all p<.01). Figure 4.9 shows performance for the three groups.

These results are consistent with the progression of performance on these scaffolds from 5th through 7th grade (Wiese & Koedinger, 2014). Like 6th graders, for the spring 5th graders addition is harder than comparison, but the scaffolds affect the difficulty of both tasks in the same way. Also, their pattern of differences in addition scores between scaffold types is closer to that of 6th graders (in the fall, all differences were significant). Finally, the comparison results were replicated with equivalence foils.
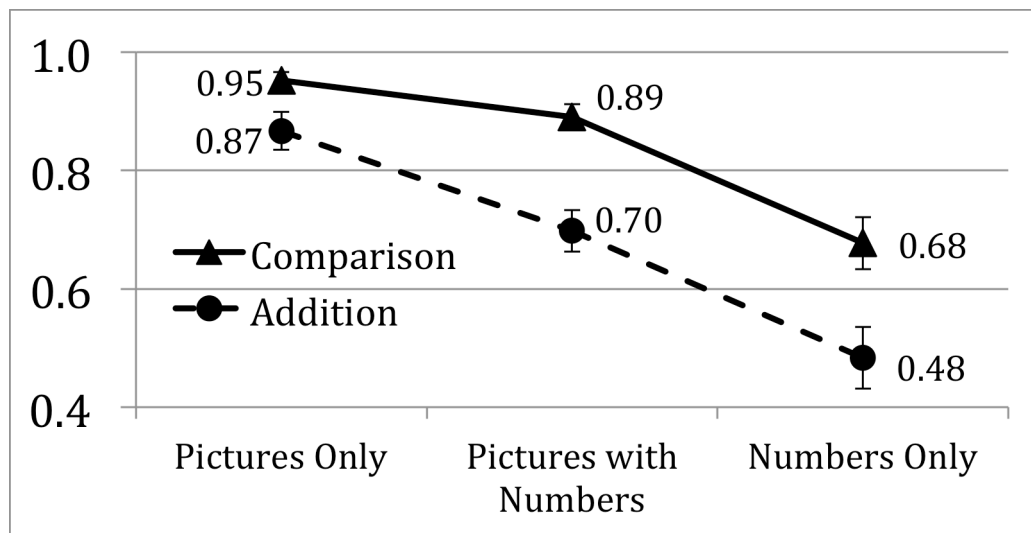
Fig. 4.9 Mean scores on the comparison and addition evaluation items by scaffold types, with bars showing standard error of the mean.

4.3.3 Error Analysis for Evaluation Items

Incorrect transfer from whole numbers is demonstrated when students say the strategy of adding both numerators and denominators (add-both) is correct, or when they say the correct answer is too big (since the numerator and denominator are both larger than the corresponding result of add-both). This error can also occur with the Pictures-Only scaffold if students count the number of shaded segments instead of comparing the overall sizes of the shaded amounts. Figure 4.10 shows the rate of correct responses and three error types (whole number error, other, and blank) for all of the addition items by scaffold type. Each student had 6 questions, and that many opportunities to make this error. Figure 4.11 shows how many whole-number errors each student made within each scaffold type. For example, 20% of students in the Pictures condition made 1 or 2 whole-number errors, while 20% of students in the Numbers condition made 5 or 6 whole-number errors. The majority of errors are consistent with whole-number thinking. These errors are most pronounced with Numbers- Only, but are mitigated by the diagrams, suggesting that the fraction symbols trigger this misconception. Together with the Addend-Sum results, mediocre performance on the Pictures with Numbers scaffolds (70%) suggests that the diagrams do not help some students tap their conceptual, qualitative understanding of addition with numbers because that qualitative understanding is not fully in place. Therefore, combining diagrams with numbers improves performance relative to numbers-only, but does not make the answers obvious for all students.
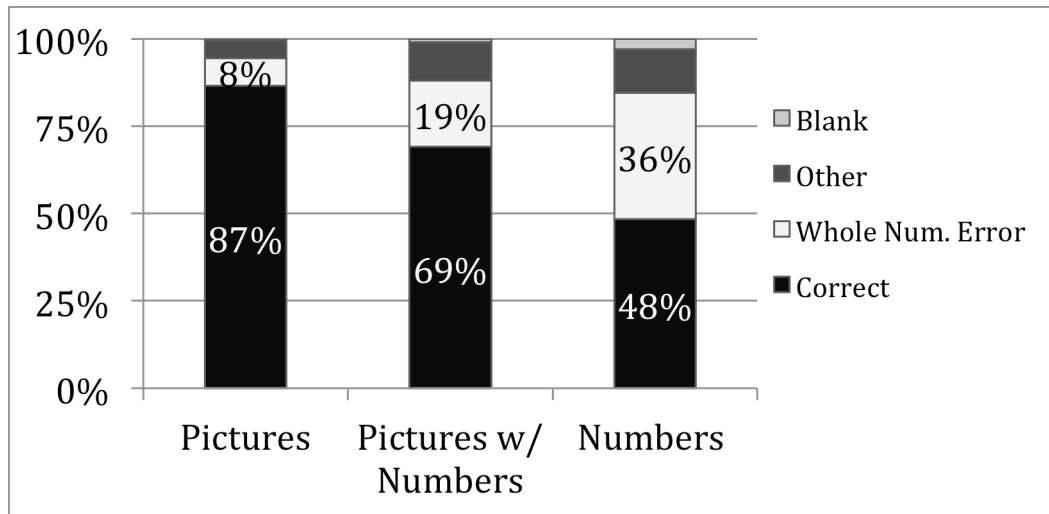
Fig. 4.10 Rates for correct answers and each possible error type, by scaffold.
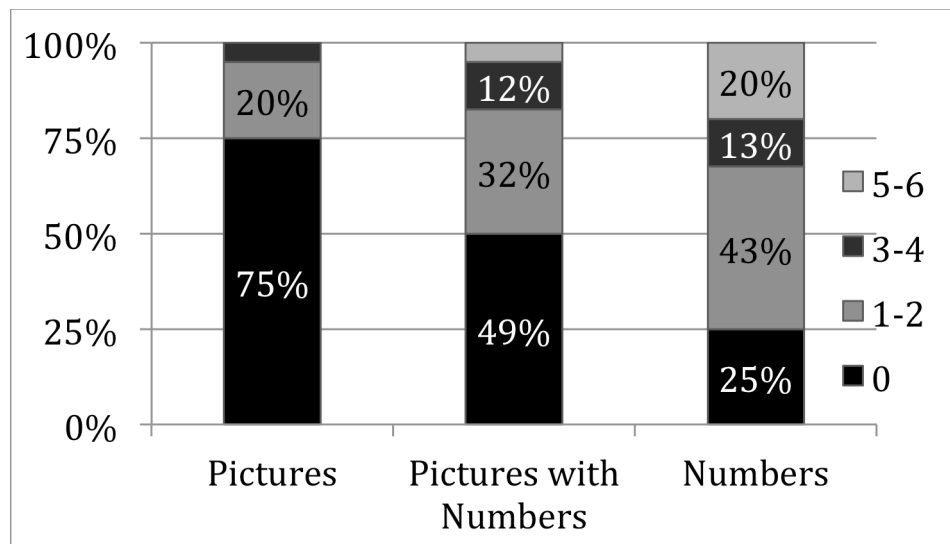

Fig. 4.11 Within each scaffold type, percentage of students who made whole-number errors at each rate.

All test forms included questions with all foil types: 3 equivalent, 5 same numerator, 2 squaring, and 2 one-less (19 students only had one). Table 4.1 shows scores by foil and scaffold type. An ANOVA on percent correct with foil type as a repeated measure and scaffold and class level as fixed factors showed a significant effect of foil (p<.01) and a significant foil by scaffold interaction (p=.023). We then ran ANOVAs for each scaffold type separately (individual

question score as dependent, foil type and class level as fixed factors and student as random factor). For Pictures Only and Pictures with Numbers, foil type was significant (p<.01) and post- hoc Tukey tests showed the One-Less foil was different from all the others (all p<.02). For Numbers Only, there was no significant effect of foil type.

	Pictures Only	Pictures with Numbers	Numbers Only
Equivalent	.97	.91	.74
Same Numerator	.98	.91	.69
Squaring	.95	.93	.61
One Less	.86	.76	.61

Table 4.1 Mean scores for the fraction comparison items, by scaffold and foil type.

Since the three scaffold types with pictures had similar results, we combine them for the error analysis. For each equivalence foil, the three error types are: mistaken for equivalent, wrong direction of inequality, and blank. Without diagrams, all four comparison items are similarly difficult, giving no evidence for consistent misconceptions. With diagrams, performance is high on all but the One-Less foil. The error analysis shows that the most popular incorrect response on the One-Less problem is that the fractions are equivalent (across the scaffold types that included pictures, 81% of responses were correct, and 15% of responses indicated that the fractions were equivalent). This error pattern is not repeated for the other foil types with pictures or for any foil type with numbers only. For numbers only, 61% of responses were correct for the One-Less foil, with 13% of responses indicating that the fractions are equivalent. With a total error rate of 39%, though the equivalence error occurs at a similar rate in terms of overall responses, it is a much lower percentage of erroneous responses (33% of erroneous responses demonstrate the equivalence error with the Numbers-Only scaffold, compared with 79% with the scaffold types that include pictures). Perhaps students who do not look closely at the pictures are fooled by the small (< 3%) size difference of the One-Less pairs. That difference is much smaller than the ~7% average difference between non-equivalent fractions in the prior study. Alternatively, perhaps students noticed the discrepancy and decided it did not matter because 1) it was close enough; and 2) adding the same number to the numerator and denominator seems similar to the correct procedure. Although pictures improved performance overall, this result is one example of their potential drawbacks: depending on their scale they may appear to show untrue relationships, and could possibly reinforce misconceptions.

4.3.4 Comparing Addends and Sums

To see if students knew that the sum of two positive, symbolic addends was bigger than each addend alone, items presented a correct addition equation and asked if the sum was bigger than each added (or visa versa). Response options were True, False, and Can't tell from the information given. Items had whole numbers, decimals, fractions, or variables (Figures 4.12-4.16). Items with variables had two presentations. This research design used a between- subjects design, assigning each student to one of the four number types, with 5 problems of that type. To control for students simply selecting true or false for all of the problems, 3 problems asked if the sum was bigger than each addend, and 2 problems asked if each addend was bigger than the sum. Students in the variables condition had 3 problems with shapes and 2 with people (Figure 4.15 and 4.16).

This addition is correct.

$$843{,}216{,}001 + 169{,}582{,}503{,}244 = 170{,}425{,}719{,}245$$

With that information only, answer these two questions:

1) 843,216,001 is bigger than 170,425,719,245

    True   False   Can't tell from the information given

2) 169,582,503,244 is bigger than 170,425,719,245

    True   False   Can't tell from the information given

Fig. 4.12 Addend-Sum item with whole numbers

This addition is correct:

$$\frac{24}{64} + \frac{39}{104} = \frac{6}{8}$$

With that information only, answer these two questions:

1) $\frac{6}{8}$ is bigger than $\frac{24}{64}$

    True   False   Can't tell from the information given

2) $\frac{6}{8}$ is bigger than $\frac{39}{104}$

    True   False   Can't tell from the information given

Fig. 4.13 Addend-Sum item with whole numbers

This addition is correct.
.617 + .083 = .7

With that information only, answer these two questions:
1) .617 is bigger than .7

    True   False   Can't tell from the information given

2) .083 is bigger than .7

    True   False   Can't tell from the information given

Fig. 4.14 Addend-Sum item with decimals

This addition is correct, but all of the numbers are covered.
All of the numbers are bigger than 0.

$$\square \; + \; \triangle \; = \; \diamond$$

With that information only, answer these two questions:
1) The number covered by ⬠ is bigger than the number covered by ☐

    True   False   Can't tell from the information given

2) The number covered by ⬠ is bigger than the number covered by △

    True   False   Can't tell from the information given

Fig. 4.15 Addend-Sum item with whole numbers

Michelle and Joe are each thinking of a number. Those numbers are both bigger than 0. When you add those numbers together, you get the number that Sandra is thinking of.

With that information only, answer these two questions:

1) Michelle's number is bigger than Sandra's number

    True   False   Can't tell from information given

2) Joe's number is bigger than Sandra's number

    True   False   Can't tell from information given

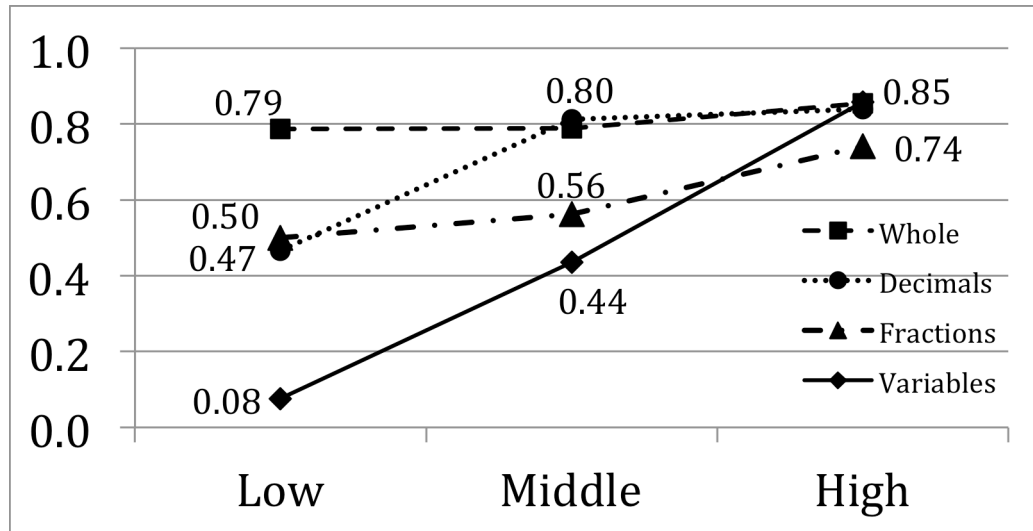**Fig. 4.16** Addend-Sum item with whole numbers

Fig. 4.17 Mean scores on Addend-Sum items by test form and class tracking level. Points differing by less than 3% were averaged.

Mean scores for each number type were 79% for whole numbers, 75% for decimals, 61% for fractions, and 51% for variables. Figure 4.17 shows performance for each number type by tracking level. I ran an ANOVA on percent correct (dependent) with test form, tracking level, and question order (forward vs. reversed) as fixed factors. Question order was not significant and there were no significant interactions with order, so I re-ran the analysis using only test form and tracking level. There was a significant effect of form, tracking level, and a significant interaction (all $p<.01$). Post-hoc Tukey tests showed significant differences between Variables and Whole Numbers and Variables and Decimals (both $p<.01$), and Fractions and Whole Numbers ($p=.022$).

Except for the High group, most students could not apply the addend-sum relationship to all four number types. This evidence supports the hypothesis that students' difficulty interpreting the fraction-addition diagrams arises from a gap in prior knowledge: they do not always recognize the significance of a proposed sum being smaller than one of the addends because they do not have a strong, fluent knowledge of the qualitative addend-sum relationship. Confusion may stem from addition with negative numbers (addition does not always make bigger) or fraction multiplication (even for positive numbers, operations do not always go in the same direction). Students can solve the whole number and decimal problems by directly comparing the numbers in each question without considering the addend-sum relationship. It is much harder to directly compare unlike-denominator fractions, and impossible for variables. This difference in strategy likely explains the significant differences between Variables and Wholes/Decimals. Performance by tracking level suggests how mastery of this relationship may develop, but appears to do so in a notation-specific way. Whole-Number performance is about the same with all three tracks, likely reflecting a direct-comparison strategy and familiarity with whole numbers. Decimals performance is low for Low- track students (~50%), likely reflecting unfamiliarity

with decimal comparison, but rises to Whole-Number level with Middle-track students. With Variables, Low-track students perform just below chance, indicating that they do not understand how addends and sums relate in the abstract. This abstract understanding trails fraction performance for Middle- and Low-track students. Although this qualitative relationship is important for reasoning about addition, these results suggest that students may not fully grasp the addend-sum relationship until they have extensive practice adding numbers of many types. This finding is in line with theories that procedural and conceptual skills develop iteratively (Rittle-Johnson, Siegler, & Alibali 2001). Further, these findings indicate that even when students can apply a concept to one symbolic context, they may not spontaneously transfer that knowledge to another symbolic context.

4.3.5 Discussion

Diagrams are thought to aid sense making by helping students apply conceptual (often qualitative) reasoning to a problem (e.g., which amounts are equal? What operation is needed?). This study provides evidence for diagrams' overall sense-making support, but also offers an explanation for why students do not always use diagrams effectively: they may lack that conceptual, qualitative reasoning that diagrams are intended to tap. This prior knowledge (e.g., that the sum of two positive addends is larger than each addend) may be obvious to adults but not to students. Further, students may be able to apply this knowledge in some contexts (e.g., with diagrams alone) but not others (e.g., the addend-sum items with fraction symbols). Still, while the knowledge that the sum of two symbolic positive addends is larger than each addend individually is related to the knowledge required for the pictures-and-numbers evaluation items, the evidence from this study suggests that it is not strictly necessary: students scored 70% on the pictures-and-numbers evaluation items and only 61% on the addend-sum items with fractions. These findings, that students do not always know how to use visual representations, support the IES Practice Guide recommendations that students be taught these skills explicitly (Woodward et al., 2012). However, the current recommended instruction focuses on mapping between a story problem and a visual representation, and then the visual representation and symbols. The results from this study suggest that students may also benefit from instruction on what type of qualitative reasoning is relevant to the problem, and how to apply that reasoning.

More generally, it seems more caution is needed in applying expert intuitions about sources of support for student sense making. While qualitative inferences may support sense making with quantitative problems, that qualitative reasoning itself may develop slowly through quantitative experience. That is, students may not apply the general relationships between addends and sums, or multiplicands and products, etc., until after they have extensive practice with those operations or equations.

5   Including Fraction Bar Pre-Instruction in the Grounded Feedback Tutor

Summary. An experiment with 163 4th and 5th graders shows improved learning with a grounded feedback tutor over a symbols-only control with step-level right/wrong feedback. Learning with grounding also transferred to symbols-only assessment items. These results hold promise for supporting representation learning in STEM domains.

5.1   Fraction Bar Pre-Instruction

To help students interpret the fraction bar representations, I added up-front instruction on the fraction bars to the grounded feedback tutor. The instruction
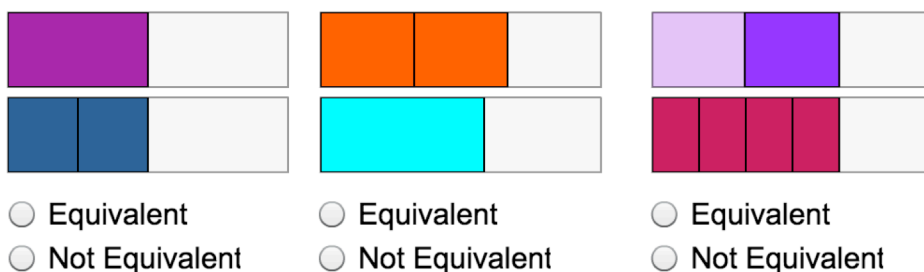

Fig. 5.1 Fraction Bar Pre-Instruction Question 1. 72% of students solved the problem, without hints, on their first try.

consists of multiple-choice problems, beginning with questions on fraction equivalence (expected to be within students' prior knowledge (Stampfer & Koedinger, 2013)) and gradually fading in the addition operations and fraction symbols. This progression is based on concreteness fading (Fyfe, McNeil, Son, & Goldstone, 2014). Students were given immediate correctness feedback and on-demand hints. Sample problems are shown in Figs. 5.1-6.

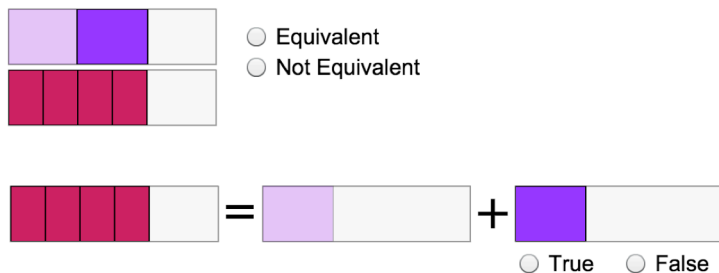Compare the rectangles, then decide if the equation is true or false.

○ Equivalent
○ Not Equivalent

○ True   ○ False

Fig. 5.2 Fraction Bar Pre-Instruction Question 5. 53% of students solved the problem, without hints, on their first try.

What does this rectangle represent?

$\frac{3}{4}$ $\frac{2}{5}+\frac{1}{10}$
○   ○

$\frac{2}{5}$ $\frac{1}{10}$

Fig. 5.3 Fraction Bar Pre-Instruction Question 10. 81% of students solved the problem, without hints, on their first try.

$\frac{1}{4}+\frac{1}{2}$

$\frac{1}{4} + \frac{1}{2} = \frac{3}{4}$
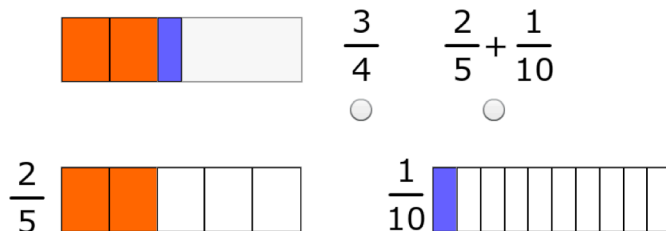
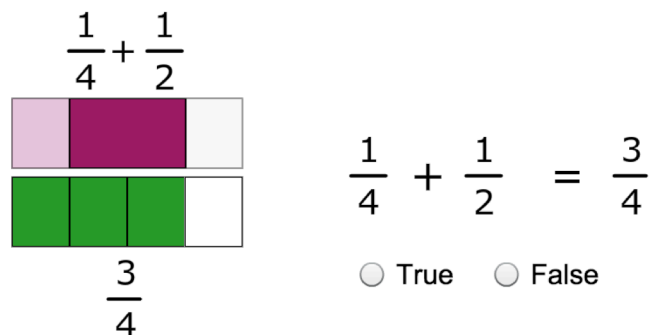$\frac{3}{4}$ ○ True   ○ False

Fig. 5.4 Fraction Bar Pre-Instruction Question 10. 81% of students solved the problem, without hints, on their first try.

5.2   Study 3: Comparing Correctness Feedback to Grounded Feedback with Pre-Instruction

This experiment compared learning with the grounded and correctness feedback tutors, using a pretest-intervention-posttest design. Both tutors included the same brief instruction on using the tutor software and on fraction addition. The grounded feedback tutor included the pre-instruction on fraction bars, described in section 2.2. This experiment investigates if pre-instruction on the feedback representation and a longer intervention time can lead to greater learning gains relative to a control. Additional research questions: (1) The grounded feedback tutor includes symbolic and graphical representations. How does learning with these representations transfer to symbolic-only contexts, and how does learning with a symbols-only representation transfer to a symbols-and-graphics context? (2) Is grounded feedback easier to work with than correctness feedback? (3) How do students leverage the grounded feedback while working with the tutor?

5.2.1   Materials, Participants, and Procedures

The 29-question pre- and posttests included 12 symbolic fraction addition items and 9 evaluation items that proposed a fraction addition equation and asked if the sum was correct, too big, or too small (3 each of pictures only, numbers only, and both pictures and numbers). Answers were scored 1 if correct and 0 otherwise. Two matched tests were counterbalanced, question order was determined randomly, and half of the tests were given in reversed question order.

194 students from 9 classes at a local public school participated in the experiment (60 4th graders and 134 5th graders). The school tracked students by achievement, and teachers identified their classes as high (3), average (5), or low (1). 31 students were removed from the sample because they were absent during the pre- or posttest, or they spent less than 45 minutes on their assigned tutor, leaving 163 students (78 grounded, 85 correctness).The experiment took place at the school during class time over four consecutive days. All random assignment was within-class. Students were given a 15-minute pretest, worked with a randomly assigned tutor for up to 80 minutes, and then took a 15-minute posttest the next day. The tests were administered on a computer and students could not return to previously answered questions.

5.2.2   Results

Did the grounded condition learn more than the correctness condition? Overall, yes. Table 5.1 shows the average scores for the overall pre- and posttests and for the three subtests, by condition. To test that pretest differences were not significant, an ANOVA was run on pretest score, with pretest order, pretest form, class tracking level, and condition as fixed factors, and class as a random factor.

The first model included all main effects and two-way interactions. After removing non-significant interactions and main effects, the final model included a marginal effect for order (p = .07), a marginal order by pretest form interaction (p = .08) and a significant class by pretest form interaction (p = .04). Condition was not significant (p = .7).

Condition	Test	Total	Addition	Evaluation	Other
Correctness	Pre	.43 (.20)	.32 (.27)	.42 (.26)	.60 (.24)
	Post	.59 (.22)	.49 (.30)	.63 (.26)	.69 (.18)
Grounded	Pre	.42 (.19)	.35 (.26)	.42 (.23)	.57 (.23)
	Post	.63 (.22)	.55 (.32)	.69 (.23)	.71 (.22)

Table 5.1. Average scores (and standard deviations) for overall tests and subtests. Paired samples t-tests show all within-condition differences from pre- to posttest are significant (p < .01)

To test if condition had a significant effect on learning, we re-ran the final model, this time on posttest score, with pretest score as a covariate. The first model included all two-way interactions with pretest score. After removing non-significant interactions and main effects, the final model included class and total pretest score as significant main effects (both p < .01) and condition as a marginal main effect (p = .065), in favor of grounded feedback. When grade was used as a fixed factor (instead of class as a random factor), the main effects model of condition, pre-test score, and grade shows a marginal effect of condition (p=.051) and a significant effect of grade (p=.016), with estimated marginal means of .564 for 4th graders, .633 for 5th graders, and .527 for the correctness condition and .624 for the grounded condition (all evaluated at a pre-test score of .426). The same tests (with class a random factor) were repeated on the addition and evaluation subtests – condition was not significant in either case.

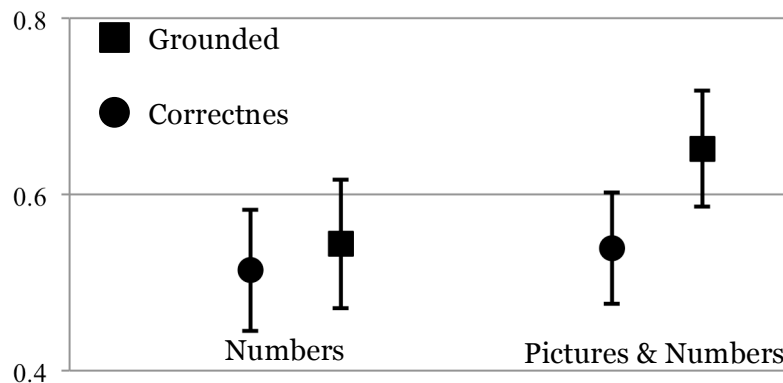How did transfer from the grounded tutor to a symbols-only assessment

Fig. 5.5 Estimated marginal means for posttest evaluation items that included numbers, with 95% confidence intervals.

88

compare to transfer from the symbols-only tutor to a dual-representation assessment? To determine if there were condition differences for scores on the numbers only and pictures and numbers evaluation items, a MANOVA was run on the posttest scores for each scaffold type, with corresponding pretest scores as covariates and class and condition as fixed factors. The condition by class interaction was not significant in the multivariate test so the model was re-run without it. Multivariate tests showed pretest scores and class were significant (p < .04), as was condition (p = .047), in favor of grounded feedback. Condition was significant on the posttest score for the pictures and numbers scaffold (p = .015, again in favor of grounding), but not for the numbers only scaffold. Figure 5.5 shows the estimated marginal means for the two scaffold types, by condition.

Since both conditions had similar gains on the addition items, greater improvement on the pictures-and-numbers evaluation items seems to explain the grounded condition's overall greater improvement from pre-test to post-test. To examine if this is the case, I ran an ANCOVA on post-test scores, excluding the addition items and the pictures-and-numbers evaluation items. With pre-test score as a covariate (also excluding the addition items and the pictures-and-numbers evaluation items), grade and condition as fixed factors, and interaction terms for grade by condition and pre-test score by condition, condition was not significant (p>.3), while pre-test score (p<.0005) and grade (p=.023) were. Neither interaction term was significant, and when the model was re-run with main effects only, the significance levels were similar: pre-test score (p<.0005) and grade (p=.022) were significant, and condition was not (p>.2). This analysis indicates that the greater improvement of the grounded condition from pre-test to post-test can be explained by the greater gains on the pictures-and-numbers evaluation items.

5.2.3 Did Students Learn from the Fraction Bar Pre-Instruction?

The fraction bar instruction aimed to help students interpret the grounded feedback. One measure of success of this instruction is how often students pressed the "done" button when the proposed sum differed from the correct sum by at least .1. Students did so on average .34 times per problem for the first 20 tutor problems and .16 times per problem (on average) for all tutor problems. Both values are much less than the .99 times per problem for Study 2 (which only included 20 tutor problems). Another measure of learning comes from a two-

Test	Correct	Whole Number Error	Other Error	Skipped
Pre	63%	30%	6%	1%
Post	63%	23%	13%	1%

Table 5.2. Proportion of correct answers and error types for the fraction bar pre-and post-test.

question pre- and posttest bracketing the pre-instruction. Similar to the question shown in Fig. 5.4, the test questions proposed a fraction addition equation with the fractions represented both symbolically and as fraction bars. Students indicated if the proposed sum was correct, too big, or too small. These pre- and posttests included one true equation and one false equation, where the sum was obtained by adding the numerators and denominators independently. Both before and after instruction, the average score was 62% correct. Errors were categorized as *whole number error*, *other error*, or *skipped*. A whole number error indicates incorrect transfer from whole number addition: answering 'correct' to a sum obtained by adding the numerators and denominators of the addends, and answering 'too big' to the correct sum. Answers that were not correct or whole number errors were coded as *other*. Table 5.2 shows the proportion of each error at the fraction bar pre- and posttest (this table includes the 95 students who completed this section, not just the 78 grounded students included in the other analyses).

After the fraction bar instruction, students had fewer whole number errors. To determine if one type of error indicates better understanding, we examined correlations between each type of error and proficiency at fraction addition problems. The study pretest included two evaluation questions that were isomorphic to those used in the fraction bar pre- and posttest, and 12 free-response symbolic fraction addition problems. For this analysis we include students who saw both of the evaluation questions, and calculated scores and error rates on the addition items based on the questions that students saw (i.e., disregarding questions that students ran out of time for). Table 5.3 shows the correlations between occurrence of each error type and (1) score on the fraction addition items and (2) rates of whole-number errors on the addition items.

Response on Fraction Addition Items	Whole Number Error	Other Error	Correct
Percent Correct	-.42*	.12	.30*
Rate of Whole-Number Error	.31*	-.11	-.21*

Table 5.3. Pearson correlations between error types on evaluation items and performance on free-response fraction addition items. *p < .03

5.2.4   Case Studies: Using Grounded Feedback

Does grounded feedback produce greater or less struggle during instruction than correctness feedback? On average, students in the grounded condition solved fewer fraction addition problems (38 vs. 74 for correctness), took longer per problem (~65 seconds per problem vs. ~40), and requested more hints per problem (1.4 vs. 0.4). These process measures show that students stuggle more when given grounded feedback. However, the difficulty inherent in engaging in

self-regulated coordination of the two reprsentations appears to be desirable (Bjork & Bjork, 2009), as in Study 2. Given that grounded feedback students learned at least as much overall, these process results suggest that grounded feedback students learn more per problem than correctness feedback students. The extra difficulties they experience are not extraneous but generative (cf. Clark & Mayer, 2011, p. 37).

How did students make use of the grounded feedback? Log data suggests two pathways: responding to the grounded feedback directly to diagnose and correct errors, and using grounded feedback to decide when to ask for a hint. Figures 5.6 and 5.7 illustrate the first strategy for a student converting 3/8 to 24ths. The student is adding 1/3 and 3/8, and got a hint for the denominator of the first fraction that said to multiply 3 by 8. The student correctly chose to multiply 8 by 3 to get the denominator for the second fraction, but then decided to multiply the
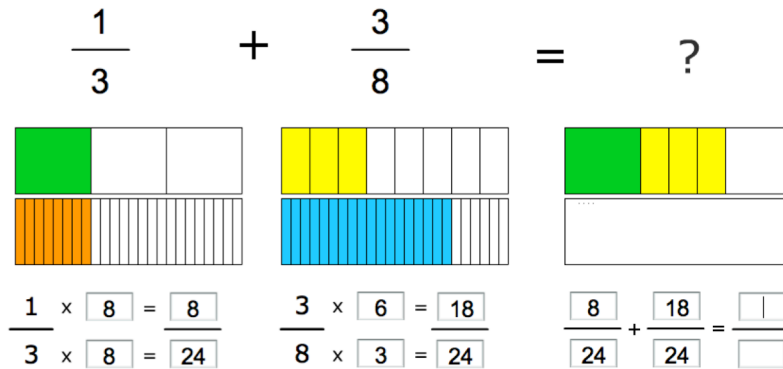

Fig. 5.6 The grounded feedback tutor. The student is converting 3/8 to 24ths.

numerator by 6. Figure 5.7 shows the student's interface at this point. The grounded feedback shows that 18/24 is bigger than 3/8. Next, the student tries 10 as a numerator (still to big), and then 9 (Fig. 5.7). After the grounded feedback shows that 9/24 equals 3/8, the student updates the multiplication area to show 3 x 3 = 9. In this case, the student does not seem able to find the equivalent fraction using symbols alone: the student does not begin by multiplying the numerator and denominator by 3. Instead, the student appears to use the grounded feedback to inform a guess-and-check strategy, identifying the direction of the error and correctly deciding when that part of the problem is complete (after converting the second fraction, the student moves on to the sum).
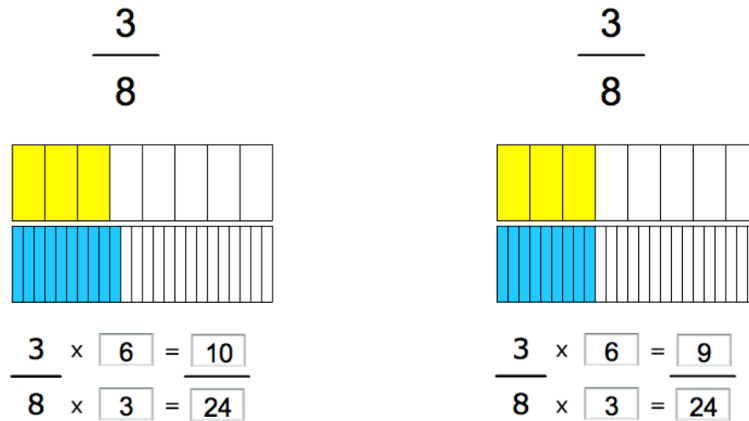
Fig. 5.7 Grounded feedback for each guess-and-check conversion attempt. After finding the correct converted fraction the student corrects the multiplication box for the numerator to indicate that 3 x 3 is 9.

In other cases, the feedback may facilitate learning from hints. In one example, a student adding 4/9 and 1/9 entered 5/18 for the sum (the *independent strategy:* adding the numerators and denominators independently). The student seems to interpret the feedback as showing an error, but appears unsure of how to fix it. Instead of pressing the done butting or guessing, the student asks for hints until the answer is provided. On the next problem, the student converts the addends incorrectly, and then uses the independent strategy on the converted fractions, again asking for a hint only after entering the incorrect sum (perhaps the student pays more attention to the addition section of the interface than the converting sections, or the student might not realize that the converted fractions should be equivalent to the addends). This student does not attempt the independent strategy on any subsequent problems. Here, the grounded feedback appears to have shaken this student's confidence in the independent strategy, perhaps facilitating acceptance of the correct strategy offered in the hints.

5.2.5 Discussion

Correctness feedback is easier to work with than grounded feedback, indicated by students solving many more correctness problems, spending less time per problem, and requesting fewer hints on each problem. How does the additional difficulty of grounded feedback affect learning? The marginal significance in favor of grounded feedback on overall learning and the non-significant difference on the addition subtest indicates that grounded feedback is no worse than correctness. The differences in learning on the evaluation items with pictures and numbers also suggest that the additional difficulties in grounded feedback are desirable. Those items include the same representations present in the grounded tutor. The numbers-only evaluation items only included the symbolic

representation present in the correctness tutor. Therefore, the pictures and numbers items can be considered target items for the grounded students while the numbers only items are transfer, and visa versa for the correctness students. With this view, the grounded feedback students were better than the correctness students at transferring their knowledge to the less-familiar format: grounded students scored just as well on the numbers only problems as the correctness students, while outperforming them on the pictures and numbers items. At the very least, the similar performance of both conditions on the fraction addition items and numbers only evaluation items shows that including the fraction bars during learning did not impede students' performance with numbers on the posttest.

Did students learn from the fraction bar tutorial? Scores on the evaluation items bracketing the pre-instruction did not change. However, students decreased their rates of whole number errors, switching to other errors instead. Whole number errors are negatively correlated with solving symbolic fraction addition problems correctly and are positively correlated with adding both numerators and denominators independently on such problems, while other errors are not correlated with either behavior. Therefore, whole number errors appear to be more harmful than other errors, and a decrease in whole number errors suggests that students benefitted from the tutorial.

These results indicate that a longer intervention time (80 vs. 40 minutes) and the inclusion of fraction bar pre-instruction addressed the shortcomings of the grounded condition in the previous study (Stampfer & Koedinger, 2012). Still, the case studies point to further possible improvements. Even with the grounded feedback, students do not always seem to recognize when their work is incorrect (e.g., a student may recognize when a proposed sum is incorrect but may not recognize when a converted fraction is incorrect). Including correctness feedback with the grounding may help: Instead of relying on the grounding alone to evaluate the action and diagnose the error, the correctness feedback will evaluate the error, freeing cognitive resources to focus on the diagnosis.

6   Comparing Grounded Feedback With and Without Correctness Feedback

Summary. An experiment with 59 4th and 5th graders compared a grounded feedback tutor to one that combined grounded and correctness feedback. Results suggest small if any differences in learning.

6.1   Motivation

Study 3 found benefits for grounded feedback over correctness feedback – while both led to similar improvement on the target addition content, grounded feedback promoted more transfer across representational contexts. While Study 2 and Study 3 compared grounded feedback as a whole to correctness feedback, the remaining studies in this thesis examine the features of grounded feedback individually. This study examines the importance of the second feature: that students evaluate their step-level work for themselves, instead of having the tutor evaluate it for them. One hypothesis is that the act of self-evaluation is crucial for learning, as it requires the student to apply their own prior knowledge to the task at hand, thus reinforcing the connections between what the student is learning and what the student already knows. An alternative hypothesis is that having immediate step-level correctness feedback will help learning by reducing cognitive load: the student is freed from the task of making the evaluation, and can focus cognitive resources on understanding *why* the action was correct or not. This study tests those competing hypotheses. Note that while this study is presented here for rhetorical reasons, it was conducted after the study presented in Chapter 7.

6.2   Study 4: Grounded Feedback With and Without Correctness Feedback

Study 4 used the grounded feedback tutor from Study 3 and compared it to a grounded plus correctness feedback tutor (described below, in section 6.2.1). Study 4 also used the same assessment forms as Study 3. Study 4 had a 10-week delay between the post-test and the delayed-test. The delayed-test happened to be scheduled on a day when the regular classroom teachers were out and when the students were sorted into gender-segregated classes to learn about their reproductive systems. Therefore, students may have been more distracted than usual on that day, and performance on the delayed-test may be less generalizable (the method for Study 4 is described in more detail in section 6.2.2).

6.2.1   The Grounded plus Correctness Tutor

The grounded plus correctness tutor uses the same basic interface as the grounded feedback tutor. Students still input symbolic fractions for both the converted addends and the sum and as they do so, corresponding fraction rectangles appear showing the entered fraction. Unlike the correctness feedback tutor (which has no fraction rectangles), students may open the addition interface, or either of the conversion interfaces, at any time. Unlike the grounded feedback tutor, students' numeric inputs are immediately colored green if correct and red if incorrect, and students may not erase correct inputs. Figure 6.1 shows a screenshot of the grounded plus correctness tutor.

Fig. 6.1 Grounded plus Correctness feedback tutor.

6.2.2 Method

Students did the study over four 40-minute class periods. On Day 1, students took a 15-minute pretest, then worked with a randomly-assigned tutor. Students continued working with the tutors on Day 2, and for the first 20 minutes of Day 3. On the second half of Day 3, students completed a 15-minute post-test. After a delay of at 10 weeks, students took a 15-minute delayed-test.

6.2.3 Participants

Two classes of 5th graders and three classes of 4th graders at the same school participated in the study. One teacher taught both of the 5th grade classes and another teacher taught all three 4th grade classes. This school was in a different district than the schools that participated in studies 2, 3, and 5. As in the previous studies, the following analyses are based on the students who completed all parts of the study, including all three assessments and at least 30 minutes of working with their assigned tutor. Table 6.1 shows the number of students, by grade and condition, who did and did not complete all parts of the study. Fisher's exact test on the students who did or did not complete the study, by condition, does not show a significant difference in attrition between the two conditions (p<.2).

Condition	Completed		Incomplete	
	5th	4th	5th	4th
Grounded	14	18	5	0
Grounded plus Correctness	15	12	4	5

Table 6.1 Number of students who did and did not complete the study, by condition. Students who did not complete the study are not included in the analysis.

6.2.4 Results: Process Measures

Students had different learning experiences with the tutors, shown by the differences in average number of problems solved, time taken per regular tutor problem, and hints requested per problem. Table 6.2 shows the means, per condition and grade, for these measures.

		Grounded	Grounded plus Correctness
Regular tutor problems attempted	5th	6.8 (2.3)	24.9 (4.6)
	4th	11.8 (2.9)	12.1 (2.1)
	All	9.5 (1.9)	19.2 (2.9)
Time taken per regular tutor problem	5th	6:08 (1:01)	2:35 (0:22)
	4th	4:43 (1:01)	4:45 (0:53)
	All	5:21 (0:47)	3:33 (0:29)
Hints per regular tutor problem	5th	14.1 (3.6)	4.5 (1.0)
	4th	8.7 (2.0)	9.8 (4.8)
	All	11.4 (2.0)	7.0 (2.3)

Table 6.2 Average number of tutor problems attempted, time taken per regular tutor problem, and hints requested per problem, with standard error of the mean in parentheses, by grade and condition.

To determine if the differences in the process measures are significant, I ran ANCOVAs on number of regular tutor problems attempted, time per regular tutor problem, and hints requested per regular tutor problem, with condition and grade as fixed factors and pre-test score as a covariate, and interaction terms for condition by pre-test score and condition by grade. For the number of regular tutor problems, the condition by pre-test interaction term was not significant so the model was re-run without it. With condition, grade, and pre-test score as main effects and a condition by grade interaction term, condition and pre-test score were significant (both $p<.0005$), as was the condition by grade interaction ($p=.042$). Grade was not significant as a main effect. Parameter estimates for the interaction term were 11.8 for 4th grade with the grounded tutor and 0 otherwise. Estimated marginal means by grade and condition are shown in table 6.3.

	Grounded	Grounded plus Correctness
5th	7.4	24.3
4th	9.9	15. 0

Table 6.3 Estimated marginal means for number of regular tutor problems attempted, by grade and condition, evaluated at a pre-test score of .24.

For the amount of time taken for each regular tutor problem, neither interaction term was significant so I re-ran the model without them. With main effects only, pre-test score was significant ($p=.003$) as was condition ($p=.022$). Grade was not significant as a main effect. Estimated marginal means were 5

minutes 31 seconds for grounded and 3 minutes 22 seconds for grounded plus correctness.

For the number of hints requested per regular tutor problem, neither interaction term was significant so I re-ran the model without them. With main effects only, pre-test score was significant (p=.019), and there was a marginal effect for condition (p=.078), with no significant effect for grade (p>.7). Estimated marginal means were 11.8 hints requested per problem in the grounded condition, and 6.5 hints requested per problem in the grounded plus correctness condition. Although the difference in the means is large, so is the 95% confidence interval for each: 7.7 – 15.9 for grounded and 2.1 – 10.8 for grounded plus correctness.

From these process measures, the grounded plus correctness tutor seems to be easier to work with than the grounded feedback tutor without correctness feedback: students in the grounded plus correctness condition solved the tutor problems more quickly, got through more problems, and requested marginally fewer hints per problem.

6.2.5   Results: Outcome Measures

To examine if there were differences between conditions at pretest, I ran an ANOVA on pre-test score with test form, order, grade, and condition as fixed factors (full factorial model). None of the interactions were significant so I re-ran the model without them. With a main-effects model, the only significant term was pre-test form (p=.017). Grade and condition were not significant (p>.2). Estimated marginal means were .285 for Form A and .195 for Form B.

Figure 6.2 shows the mean scores for each assessment, by condition. Figure 6.3 shows the mean scores on the addition items for each assessment, by condition.
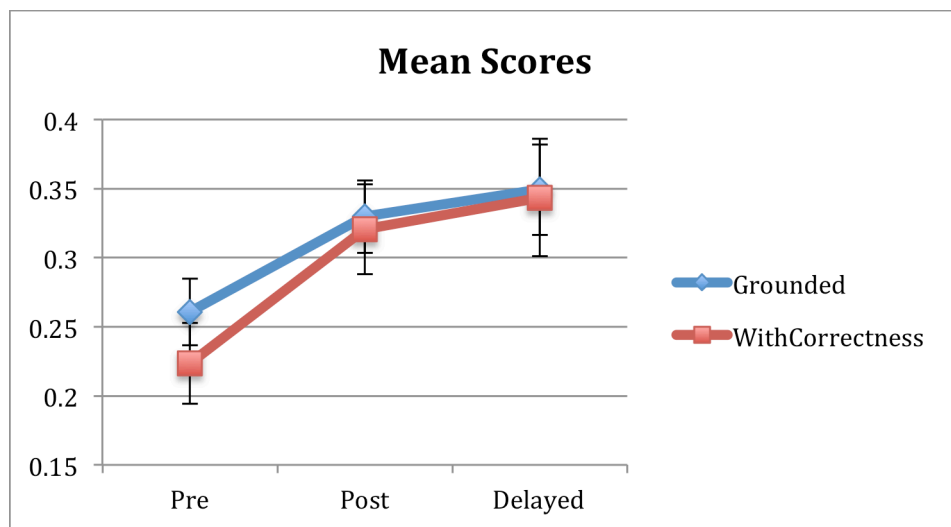

Figure 6.2 Mean score at each assessment time, by condition, with bars showing the standard error of the mean.
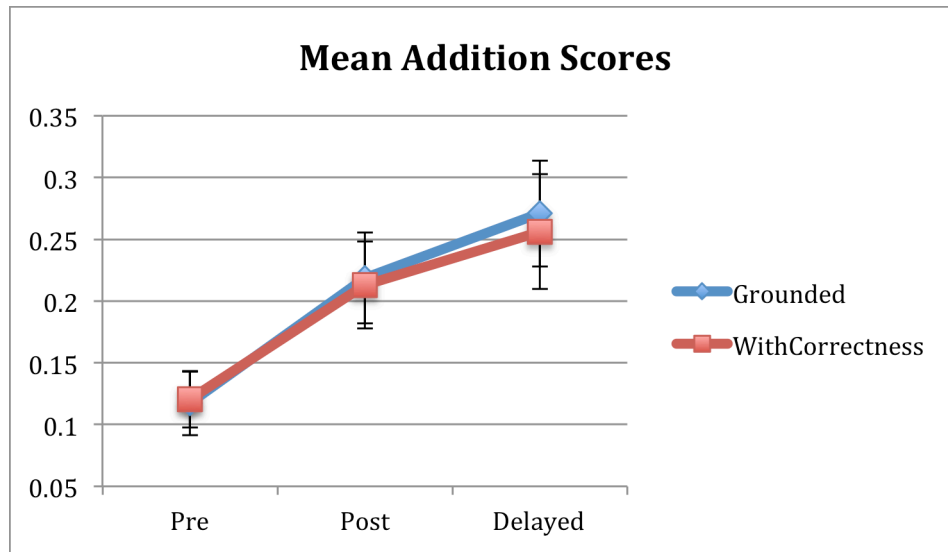
Figure 6.3 Mean addition scores at each assessment time, by condition, with bars showing the standard error of the mean.

Overall Learning To determine if immediate learning differed by condition, I ran an ANCOVA on the post-test scores, with grade, condition, and pre-test form as fixed factors and pre-test score as a covariate (pre-test form is included since it was a significant effect for pre-test score). With a full-factorial model and a condition by pre-test score interaction term, there is a significant effect of the three-way interaction between condition, grade, and pre-test form (p=.04). Re-running the model with that interaction and the main effects (since the other interactions were not significant), pre-test score is significant (p<.0005) and the three-way interaction is marginal (p=.054), with no other significant effects. Estimated marginal means for post-test scores, by condition, grade, and pre-test form, are shown in table 6.4.

		Pretest Form A	Pretest Form B
Grounded	5th	.356	.256
	4th	.287	.379
Grounded plus Correctness	5th	.317	.421
	4th	.268	.306

Table 6.4 Estimated marginal means for post-test score, by grade, condition, and pre-test form, evaluated at a pre-test score of .243.

To determine if there were differences in retention and future learning by condition, I ran an ANCOVA on the delayed-test scores, with grade, condition, and pre-test form as fixed factors and post-test score as a covariate. Note that pre-test form and delayed-test form are the same. With a full-factorial model and

a condition by post-test score interaction term, none of the interaction terms were significant, so I re-ran the model without them. With main effects only, post-test score was significant (p<.0005), pre-test form was marginal (p=.066) and there were no significant effects for grade or condition (both p>.1). Estimated marginal means for delayed-test score are .378 for delayed-test form A and .318 for delayed-test form B.

To determine if there were differences in learning across the entire study, I ran an ANCOVA on the delayed-test scores, with grade, condition, and pre-test form as fixed factors and pre-test score as a covariate. While the two prior analyses compared scores on different test forms, this analysis compares scores on the same test form, and therefore may more reliable, given the significant effect of test form in those analyses. With a full-factorial model and a condition by pre-test interaction term, none of the interactions were significant so I re-ran the model without them. With main effects only, pre-test form was not significant, so I re-ran the model with condition, grade, and pre-test score as main effects. With the new model, pre-test score was significant (p<.0005) while grade and condition were not (p>.4). Overall, these results do not provide evidence supporting the hypothesis that condition had a significant effect on students' learning.

Addition Learning To examine if there were differences between conditions on the target addition items at pretest, I ran an ANOVA on pre-test addition score with test form, order, grade, and condition as fixed factors (full factorial model). None of the interactions were significant so I re-ran the model without them. With a main-effects model, there was a marginal effect for grade (p=.091) and test order (p=.052), with no significant effect for condition (p>.9) or pre-test form (p>.1). Surprisingly, the 4th graders had higher pre-test addition scores than the 5th graders: estimated marginal means were .142 for the 4th graders and .084 for the 5th graders.

To determine if there were differences in immediate learning for the target addition content, I ran an ANCOVA on post-test addition scores, with condition, grade, and test order as fixed factors and pre-test addition score as a covariate. Test order is included since it was significant for pre-test scores. With a full-factorial model, including a condition by pre-test form interaction term, none of the interactions were significant. With a main-effects model, pre-test addition score was significant (p<.0005), while condition, grade, and pre-test order were not (all p>.6).

To determine if there were differences in retention and future learning for the target addition content, I ran an ANCOVA on delayed-test addition scores, with condition, grade, and test order as fixed factors and post-test addition score as a covariate. With a full-factorial model, including a post-test addition score by condition interaction term, none of the interactions were significant so the model was re-run without them. With main effects only, post-test addition score was significant, while condition, grade, and test order were not (all p>.4).

To determine if there were differences learning for the target addition content across the whole study, I ran an ANCOVA on delayed-test addition scores, with condition, grade, and test order as fixed factors and pre-test addition score as a covariate. With a full-factorial model, including a pre-test addition score by condition interaction term, none of the interactions were significant so the model was re-run without them. With main effects only, pre-test addition score was significant, while condition, grade, and test order were not (all p>.4). As with the scores on the test overall, these results do not provide evidence supporting the hypothesis that condition affected learning on the target addition content.

Evaluation learning. Mean scores on the evaluation items are show in figure 6.4. To determine if there were differences on the evaluation items at pre-test, I ran a MANOVA on the pre-test evaluation scores for each scaffold type (rectangles, numbers, and both), with test form, order, grade, and condition as fixed factors. Multiovariate tests show a significant effect for the three-way interaction of condition, pre-test form, and order, but none of the other interactions were significant. Re-running the model with main effects and the three-way interaction of condition, pre-test form, and order, the interaction was marginal (p=.051, Pillai's Trace) and there was a significant effect of order (p=.038) with no other significant effects. Between-subject effects show that the interaction term and the order term are only significant for the rectangles and numbers scaffold (both p<.03). Since order was significant at pre-test it will be included in the analyses of learning.

To determine if there were differences in immediate learning by condition, I ran a MANOVA on the post-test evaluation scores for each scaffold type with the pre-test evaluation scores for each scaffold type as covariates, and condition, grade, and order as fixed factors. With a full-factorial model, none of the interactions were significant so I re-ran the model without them. With a main effects model, the only significant effect was the pre-test score for the numbers-only scaffold (p=.005), with a marginal effect for order (p=.08). Tests of between-subject effects show that the pre-test numbers-only evaluation score is significant for post-test scores for the numbers-only scaffold (p=.001) and the numbers and rectangles scaffold (p=.026), but not the rectangles-only scaffold (p>.8). Order is significant only for the post-test evaluation scores for the numbers-only scaffold (p=.017, with parameter estimates showing a benefit for the forward order).

To determine if there were differences in learning by condition across the whole study, I ran a MANOVA on the delayed-test evaluation scores for each scaffold type with the pre-test evaluation scores for each scaffold type as covariates, and condition, grade, and order as fixed factors. With a full-factorial model, none of the interactions were significant so I re-ran the model without them. With a main effects model, multivariate tests show that the pre-test evaluation score for the rectangle scaffold was marginal (p=.052) and condition was significant (p=.012), with no other significant effects. Tests of between-subject effects show that condition is significant only for the delayed-test evaluation scores for the
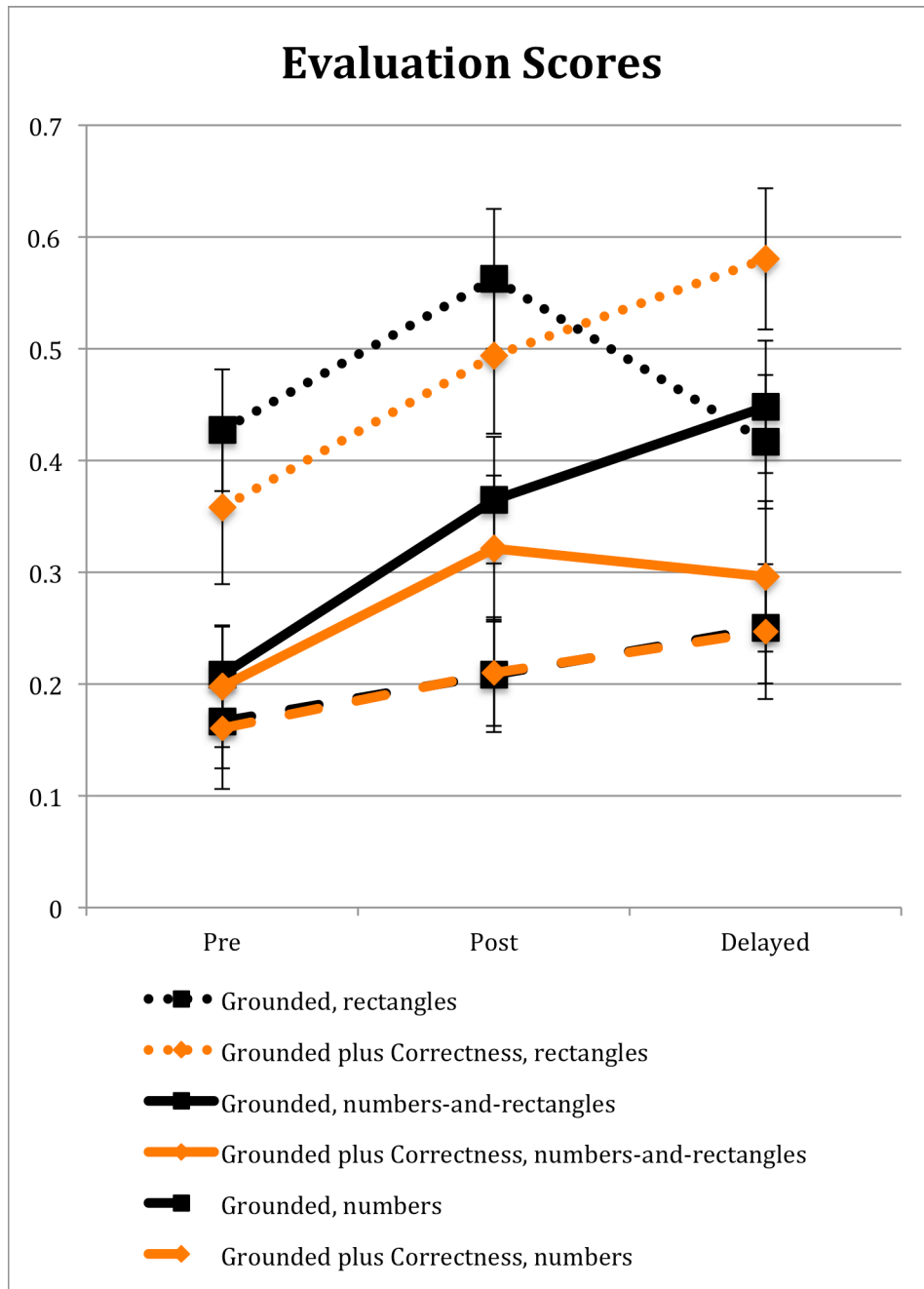
Figure 6.4 Mean evaluation scores at each assessment time, by condition and scaffold, with bars showing the standard error of the mean. The orange lines with diamond icons show scores for the Grounded condition, and the black lines with square icons show scores for the Grounded plus Correctness condition. Differences between the conditions for the numbers-only scaffold are hard to distinguish because the points overlap.

rectangle scaffold (p=.04), and that pre-test evaluation scores for the rectangle scaffold are significant for both the delayed-test rectangle scaffold (p=.048) and the delayed-test numbers and rectangles scaffold (p=.012). Parameter estimates show a benefit for the grounded plus correctness condition on the delayed-test rectangles-only evaluation scores (B = -.184 for grounded).

These results suggest a benefit for the grounded plus correctness condition for improvement on the rectangles-only evaluation task across the study as a whole.

6.2.6 Discussion and Conclusion

This study compared grounded feedback to grounded plus correctness feedback. The grounded plus correctness tutor was easier for students to use: they needed less time per fraction addition problem, and requested fewer hints per problem. However, these differences in process measures did not lead to significant differences in learning outcomes: there were no significant differences in students' learning as measured by the full assessments or by the addition items alone. On the evaluation items, grounded plus correctness improved more on the rectangle-only evaluation items from the pre-test to the delayed-test. This result suggests that grounded plus correctness may be slightly better than grounded alone. However, this finding should be interpreted cautiously as it was not based on a prior hypothesis, but rather an exploratory data analysis.

Study 4 tested the importance of the second feature of grounded feedback: that students evaluate their step-level work on their own. Since the results did not show significant differences in overall learning for students in the grounded plus correctness condition compared with the grounded condition, they do not provide evidence that the second feature is important for student learning.

7 Comparing Correctness Feedback, Virtual Manipulatives, and the Combination of Grounded and Correctness Feedback

Summary. An experiment with 191 4th and 5th graders compared three tutoring conditions: correctness feedback, a virtual manipulatives tutor, and tutor that provided both grounded and correctness feedback. Results indicate relative benefits for immediate fraction addition learning for the correctness condition, with no significant differences in addition learning across the whole study. The two conditions that included fraction bars (the combination grounded plus correctness feedback tutor and the virtual manipulatives tutor) showed benefits in particular for students with lower pre-test scores.

7.1 Motivation

This chapter further investigates the importance of the individual features of grounded feedback. The three-condition study presented in this chapter compares the correctness feedback control to a tutor that combines grounded and correctness feedback and to a virtual manipulatives tutor that provides the same visual information as grounded feedback, but has students act on the concrete representation instead of the abstract one. Theories of cognitive load, split

attention, and the affordances of concrete and abstract representations lead to conflicting hypotheses for the relative benefits of these conditions.

7.1.1 Hypotheses

We contrast two pairs of conflicting hypotheses:

1a) Grounded plus correctness feedback will lead to more robust learning than correctness feedback. While students benefitted from grounded feedback alone, correctness feedback will provide additional support and will prevent unproductive floundering by ensuring that students do not erase correct inputs. Correctness feedback may also help students take better advantage of the grounded feedback. Correctness feedback will clearly communicate if a students' action was right or wrong, and may then prompt the student to carefully consider the grounded feedback to decide why that action makes sense or not. Since studies 2 and 3 found benefits for grounded feedback over correctness feedback, enhancing grounded feedback will only widen that difference.

1b) Grounded plus correctness feedback will lead to poorer learning than correctness feedback alone. Since the correctness feedback is easier to interpret than the grounded feedback, students will have no incentive to spend time and effort interpreting the fraction rectangles. The dynamic fraction rectangles will simply be distracting, and will lead to split attention. Without paying full attention to the correctness feedback or the grounded feedback, students will be worse off than if they were using correctness feedback alone.

2a) Virtual manipulatives will serve the same function as grounded feedback, and therefore will lead to more robust learning than correctness feedback alone. While having students act on the fraction rectangles instead of on the numeric symbols will change how students interact with the tutor, it will not change how students think about the information that is presented.

2b) Virtual manipulatives will lead to poorer learning than symbols-only correctness feedback. Since students will directly manipulate the concrete representation, they will ignore the harder-to-interpret symbolic representation, and students' learning with the concrete representation will not transfer easily to symbols-only contexts. In this view, changing the input mode from the abstract representation to the concrete representation will change not only how students interact with the system but also which interface elements they attend to.

7.2   Materials

The correctness feedback tutor, assessments, and rectangle pre-instruction were identical to those used in studies 3 and 4. The grounded plus correctness tutor was the same as the one used in study 4. The virtual manipluatives tutor has the same basic structure as the grounded Feedback tutor. Both tutors provide on-demand text hints and ensure that the current problem is solved correctly before allowing the student to move on to the next one.

7.2.1   The Virtual Manipulatives Tutor

The virtual manipulatives tutor uses the same basic interface as the grounded feedback tutor, except that students do not enter numbers for the converted and sum fractions. Instead, students act on the fraction bars (figure 7.1). The fraction

Fig. 7.1 Virtual Manipluatives tutor.

bars for the converted and sum fractions are controlled by a set of buttons directly beneath them. The 'Denominator' buttons determine how many equal pieces the rectangle is divided into, and the 'Numerator' buttons determine how many of those pieces are colored in. The "+1" and "-1" buttons change one piece at a time, and the "+5" and "-5" buttons change 5 pieces at a time, to facilitate the construction of fractions with large numerators and denominators (e.g., if the student starts with the fraction 3/9, pressing the "-1" button for the numerator and the "+5" button for the denominator will yield the fraction 2/14). Alternatively, students may click on a piece of a fraction bar to color or un-color it. When students use the numerator buttons, pieces are colored in from left to

right and un-colored from right to left. When students click on a piece to color or un-color it, they may do so in any order. Students' actions on the fraction bars are reflected with the numeric fractions below. As in the grounded feedback tutor, the numbers that students multiply with to find the converted fractions are not connected to the fraction bars, therefore, in the virtual manipulative tutor students enter those symbolic numbers. Like the grounded feedback tutor, the virtual manipulatives tutor does not provide immediate correctness feedback, and allows students to open the conversion and addition interfaces at any time.

   As the sheer number of interface elements in the virtual manipulatives tutor might be overwhelming, and as students may not realize that they are intended to use the buttons to control the fraction bars, students are introduced to the button interface with a brief tutorial. The tutorial starts with a modified conversion interface, with just the "+1" and "-1" buttons for the denominator. Students are instructed to divide a fraction bar into 4 pieces and color in one of the pieces by clicking on it (figure 7.2). Next, the "+5" and "-5" buttons for the denominator are introduced, to show students that they can change multiple pieces at a time (figure 7.3). Finally, students are shown the full set of buttons for the numerator and the denominator (figure 7.4).

Fig. 7.2 Virtual Manipluatives tutorial 1.

Fig. 7.3 Virtual Manipluatives tutorial 2.

Fig. 7.4 Virtual Manipluatives tutorial 3.

7.3 Study 5: Comparing Three Conditions

Like Studies 2 and 3, Study 5 took place in a school district near Pittsburgh, and students completed all study activities in school during normal class time. Like Studies 2 and 4, Study 5 included a delayed post-test.

7.3.1 Participants and Method

4th and 5th graders from the same school district participated in this study. All of the 5th graders attended the same school (6 classes), while the 4th graders attended two different schools (3 classes at one school, 2 in the other). Like the previous studies, this study used within-class random assignment. Students did the study according to the same schedule, but each school started on different day.

Students did the study over 40-minute class periods. On Day 1, students took a 15-minute pretest, then worked with a randomly-assigned tutor. Students continued working with the tutors on Day 2, and for the first 20 minutes of Day 3. On the second half of Day 3, students completed a 15-minute post-test. After a delay of at least 2.5 weeks, students took a 15-minute delayed-test. Due to scheduling constraints, the delay was not the same for all classes. Students used the same A and B test forms as in Study 3. Students were given one test form as the pre-test and delayed-test, and the other test form as the post-test. Students were randomly assigned to see the items in either forward or reversed order across all three tests. Students in the two conditions with fraction bars were given the fraction bar tutorial, as in Study 3. Students in the virtual manipluatives condition also did the virtual manipulatives tutorial (most students did the tutorial before starting any of the regular tutor problems; some students who started the regular tutor problems on Day 1 did the tutorial on Day 2).

An adjustment to the study schedule was made for three of the six 5th grade classes. Two pairs of classes were originally scheduled to do the study at the same time. On Day 2 of the study, the tutors ran extremely slowly when the first pair of classes used them at the same time. I scheduled a make-up day with these classes, so they had an extra class period to work with their tutors. Instead of three consecutive study days, students in those two classes did Day 1, Day 2, and a make-up day consecutively. After a gap of two weekend days and two school days, students did Day 3 (20 minutes with the tutors and the 15-minute post-test). I also rescheduled one of the classes from the second pair: instead of three consecutive study days, the did Day 1, a gap of one school day, Day 2, a gap of two weekend days and two school days, and then Day 3 (on the same day as the other two re-scheduled classes). The other class in the second pair did the study as originally scheduled.

244 students began the study. Some students did not complete all three assessments or spent less than 30 minutes working with their assigned tutor, leaving 191 students in the analysis. A chi-square test on the number of students who did or did not complete the study, per condition, indicates that attrition did not differ by condition (p>.3). Table 7.1 shows the number of students who did and did not complete the study, by grade and condition.

Condition	Completed		Incomplete	
	5th	4th	5th	4th
Correctness	41	25	9	4
Grounded plus Correctness	33	26	13	6
Virtual Manipulatives	40	26	13	8

Table 7.1 Number of students who did and did not complete the study, by condition. Students who did not complete the study are not included in the analysis.

7.3.2 Process measures: Results and Analysis

Students had different learning experiences with the tutors, shown by the differences in average number of problems solved, time taken per regular tutor problem, and hints requested per problem. Table 7.2 shows the means, per condition, for these measures.

	Correctness	Grounded plus Correctness	Virtual Manipulatives
Regular tutor problems attempted	54 (3.8)	21 (1.8)	8.5 (.88)
Time taken per regular tutor problem	1:33 (0:10)	3:57 (0:21)	9:37 (0:38)
Hints per regular tutor problem	2.6 (.56)	4.8 (.77)	16.3 (1.6)

Table 7.2 Average number of tutor problems attempted, time taken per regular tutor problem, and hints requested per problem, with standard error of the mean in parentheses.

To determine if the differences in the process measures are significant, I ran ANCOVAs on number of regular tutor problems attempted, time per regular tutor problem, and hints requested per regular tutor problem, with condition and grade as fixed factors and pre-test score as a covariate, and interaction terms for condition by pre-test score and condition by grade. For the number of regular tutor problems, the condition by grade interaction term was not significant so the model was re-run without it. Grade was significant (p=.001), as was pre-test score (p<.0005), and the condition by pre-test interaction (p<.0005). Condition was not significant as a main effect. Parameter estimates for the pre-test by

condition interaction term were 124 for Correctness (significant at 0<.0005) and 26 for grounded plus correctness (p>.1). Estimated marginal means for regular tutor problems attempted were 54 for correctness, 22 for grounded plus correctness, and 9 for virtual manipulatives (evaluated at a pre-test score of .28). Pairwise comparisons show that the differences between all three conditions are significant (p<.0005 for all comparisons) with a Bonferroni correction for multiple comparisons.

For time taken per regular tutor problem, the interaction terms for condition by grade and condition by pre-test score were not significant so the model was re-run without them. With main effects only, condition and pre-test were significant (both p<.0005), and grade was not (p>.5). Estimated marginal means for each condition were 1 minute 35 seconds for correctness, 3 minutes 46 seconds for grounded plus correctness, and 9 minutes 41 seconds for virtual manipulatives (evaluated at a pre-test score of .28). Pairwise comparisons show that the differences between all three conditions are significant (p=.001 for correctness and grounded plus correctness, p<.0005 for the other comparisons) with a Bonferroni correction for multiple comparisons.

For number of hints per regular tutor problem, the interaction terms for condition by grade and condition by pre-test score were not significant so the model was re-run without them. With main effects only, condition was significant (p<.0005), as was grade (p=.005), and pre-test score (p=.007). Parameter estimates by condition were 3.2 hints per problem for correctness, 4.8 for grounded plus correctness, and 16.8 for virtual manipulatives (evaluated at a pre-test score of .28). Pairwise comparisons show that the differences between virtual manipulatives and the other two conditions are significant (p<.0005), while the difference between correctness and grounded plus correctness is not significant (p>.8), with a Bonferroni correction for multiple comparisons.

Engagement with Numbers in the Virtual Manipulatives Condition. One hypothesis predicted that the interface design of the virtual manipulatives tutor would draw students' attention away from the fraction symbols, and therefore students would learn procedures that worked easily on fraction bars, but would not transfer redily to fraction symbols. However, students did not always interact with the tutor as predicted. While students could not type into the converted or sum fraction input areas once they had manipulated the fraction bars, a bug in the tutor design allowed students to type in numbers beforehand. These numbers would be overwritten when input was provided to the fraction bars. 90% of students in the Virtual Manipulatives condition inputted numbers in this manner while working with the tutor. On average, 50% of a student's attempted problems invovled numeric input. This behavior suggests that the interface design did not draw students' focus away from the numeric sumbols.

7.3.3 Outcome Measures for the Three Conditions

To examine if there were differences between conditions at pretest, I ran an ANOVA on pre-test score with test form, order, grade, and condition as fixed factors (full factorial model). Grade was significant (p<.0005), as was a grade by order interaction (p=.027; forward order was easier for 4th graders, with a parameter estimate of .103). There were not significant pre-test differences by
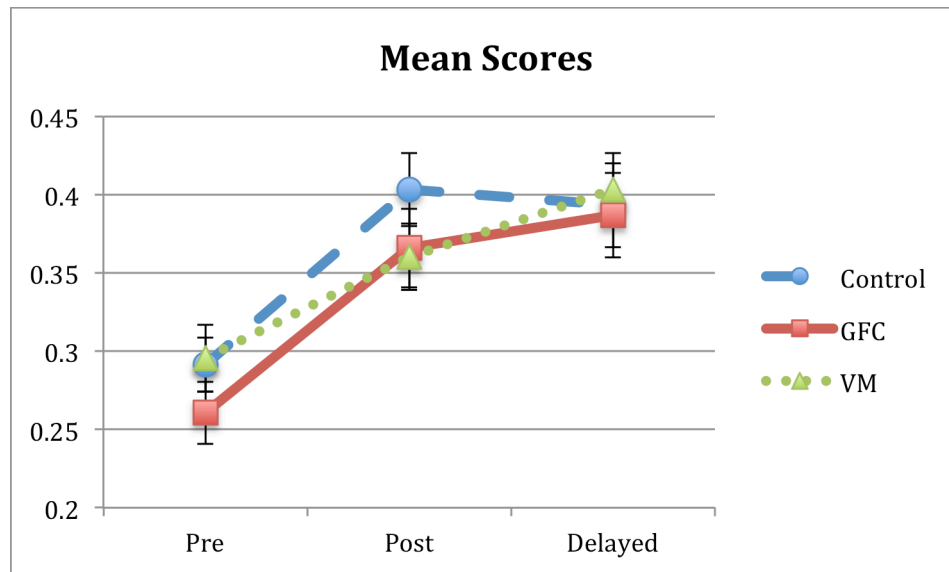

Figure 7.5 Mean score at each assessment time, by condition, with bars showing the standard error of the mean.
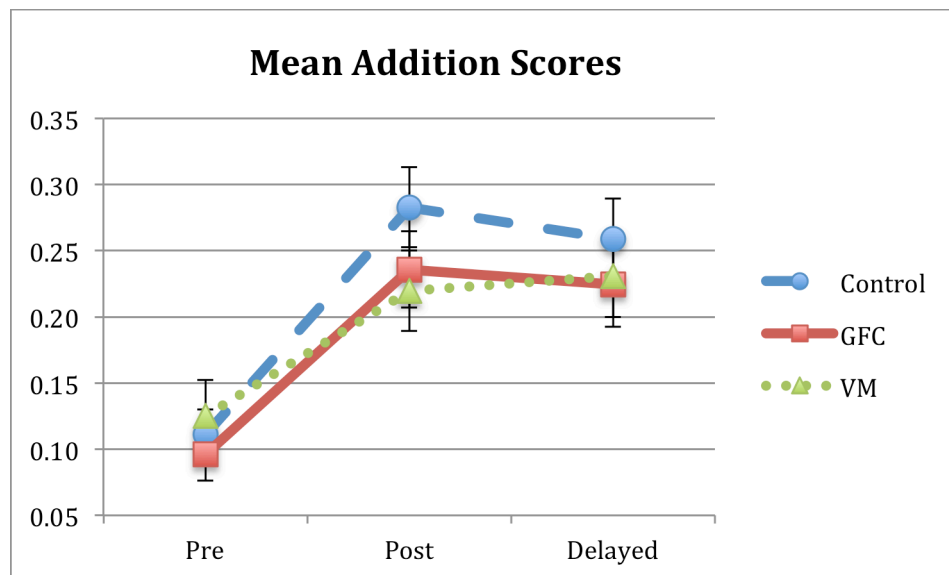

Figure 7.6 Mean scores for the addition items, at each assessment time, by condition, with bars showing the standard error of the mean.

condition. Figure 7.5 shows the mean scores for each assessment, by condition. Figure 7.6 shows the mean scores on the addition items for each assessment, by condition.

Overall Learning. To determine if condition led to differences in immediate learning, I ran an ANCOVA on post-test scores with condition, grade, and order as fixed factors (since there was a significant interaction of grade and test form order at pretest), with a full factorial model, including a condition by pre-test score interaction term. None of the interactions were significant, so I re-ran the model without them. With main effects only, there was a significant effect of pre-test (p<.0005) and a marginal effect of grade (p=.06). There were no significant differences in pre-to-post learning by condition. Since test form order was not significant, I re-ran the main effects model without it. With condition, grade, and pre-test score, pre-test remained significant (p<.0005), grade remained marginal (p=.056), and condition remained not significant (p>.1). Since test form order was not significant for immediate learning, it is not included in further models.

To determine of condition led to differences in retention and future learning, I ran an ANCOVA on delayed-test scores, with condition and grade as fixed factors, post-test score as a covariate, and interaction terms for condition by grade and condition by post-test score. None of the interactions were significant so I re-ran the model without them. With a main-effects model, post-test score is significant (p<.0005), grade is not significant (p>.1), and condition is marginally significant (p=.078). Estimated marginal means show a delayed-test score of .367 for correctness, .395 for grounded plus correctness, and .415 for virtual manipulatives (evaluated at a post-test score of .377). Pairwise comparisons indicate that the difference in learning between correctness and virtual manipulatives is marginally significant (p=.074), and all other pairwise comparisons are not significant (using the Bonferroni correction to adjust for multiple comparisons).

To determine if condition led to differences in learning across the whole study, I ran an ANCOVA on delayed-test scores, with condition and grade as fixed factors, pre-test as a covariate, and interaction terms for condition by grade and condition by pre-test score. There was a significant effect for condition by pre-test score (p=.013) but not for condition by grade (p>.8) so the model was re-run without the latter. With the new model, condition was significant (p=.019), as was pre-test (p<.0005), and the interaction between condition and pre-test (p=.011), with grade marginal (p=.06). Parameter estimates for the interaction term are .523 for Correctness and .198 for grounded plus correctness. Estimated marginal means for delayed-test are .379 for correctness, .402 for grounded plus correctness, and .391 for virtual manipulatives (evaluated at a pre-test score of .28). None of the pairwise comparisons for condition were significant (with the Bonferroni correction for multiple comparisons; condition is not significant if the pre-test by condition interaction term is not included in the model).

Addition Learning. To determine if there were condition differences on the addition item scores at pre-test, I ran an ANOVA on the pre-test addition scores, with test form, order, grade, and condition as fixed factors. Grade was significant (p<.0005) and there was a marginal grade by order interaction (p=.074). Order will be included in subsequent analyses for addition scores.

To determine if there were differences by condition for immediate learning of the target addition content, I ran an ANCOVA on the post-test addition scores, with condition, grade, and order as fixed factors and pre-test addition score as a covariate. With a full-factorial model and a condition by pre-test addition score interaction term, none of the interactions were significant, so the model was re-run without them. Order was not significant as a main effect in either model, so the model was re-run without it. All terms in the main-effects model were significant (p<.0005 for pre-test addition score; p=.002 for grade; and p=.049 for condition). Estimated marginal means for post-test addition score by condition are .273 for correctness, .243 for grounded plus correctness, and .199 for virtual manipulatives (evaluated at a pre-test addition score of .11). Pairwise comparisons show that the difference between correctness and virtual manipulatives is significant (p=.045) and the other differences are not (with the Bonferroni correction for multiple comparisons).

To determine if there were differences by condition for retention and future learning of the target addition content, I ran an ANCOVA on the delayed-test addition scores, with condition and grade as fixed factors and post-test addition score as a covariate. In a full-factorial model with an interaction term for condition by post-test addition score, neither of the interactions were significant so the model was re-run without them. With main effects only, condition and grade were not significant (p>.3) and post-test addition score was significant (p<.0005).

To determine if there were condition differences across the whole study on the target addition content, I ran an ANCOVA on the delayed-test addition scores with condition and grade as fixed factors and pre-test addition score as a covariate. In a full-factorial model with a condition by pre-test interaction term, none of the interactions were significant so I re-ran the model without them. With main effects only, pre-test addition score was significant (p<.0005) as was grade (p=.018); condition was not significant (p>.4).

7.3.4 Collapsing the Two Fraction Bar Conditions

In this set of analyses, the grounded plus correctness condition and the virtual manipulatives are collapsed, to determine if the presence of fraction bars affects learning (compared to the symbols-only control). Differences at pre-test were assessed with an ANOVA on pre-test score, with condition (with or without rectangles), pre-test form, order, and grade as fixed factors. With a full-factorial model, there was a marginal effect for the grade by order interaction (p=.052) and the three-way interaction of grade, pre-test form, and pre-test order (p=.09). Re-running the model without the non-significant interactions, there was a

significant effect for grade (p<.0005) and a marginal grade by order interaction (.092), with no significant differences by condition (p>.5), pre-test form (p>.7) or order as a main effect (p>.3).

Overall Learning. To determine if condition led to differences in immediate learning, I ran an ANCOVA on post-test scores with condition (with or without rectangles), grade, and order as fixed factors (since there was a marginal interaction of grade and test form order at pretest), with a full factorial model, including a condition by pre-test score interaction term. There was a marginal effect for the grade by condition interaction (p=.097) but not for the other interaction terms, I re-ran the model without them. With main effects and the grade by condition interaction, test form order was not significant, so I re-ran the main model without it. With the main effects of grade, condition, pre-test score, and a grade by condition interaction term, grade was significant (p=.017), as was pre-test score (p<.0005), with a marginal grade by condition interaction (p=.065). Condition was not significant as a main effect (p>.3). Parameter estimates for the condition by grade interaction are -.08 for 4th grade with correctness and 0 for all other combinations. Since test form order was not significant for immediate learning, it is not included in further models.

To determine of condition led to differences in retention and future learning, I ran an ANCOVA on delayed-test scores, with condition and grade as fixed factors, post-test score as a covariate, and interaction terms for condition by grade and condition by post-test score. Neither of the interactions were significant so I re-ran the model without them. With a main-effects model, post-test score is significant (p<.0005), as is condition (p=.039), with no significant effect for grade (p>.1). Estimated marginal means show a delayed-test score of .367 for correctness and .405 for rectangles (evaluated at a post-test score of .377).

To determine if condition led to differences in learning across the whole study, I ran an ANCOVA on delayed-test scores, with condition and grade as fixed factors, pre-test as a covariate, and interaction terms for condition by grade and condition by pre-test score. There was a significant effect for condition by pre-test score (p=.006) but not for condition by grade (p>.7) so the model was re-run without the latter. With the new model, condition was significant (p=.006), as was pre-test (p<.0005), and the interaction between condition and pre-test (p=.005), with grade marginal (p=.066). The parameter estimate for the interaction of correctness and pre-test score was .446. Estimated marginal means for delayed-test are .379 for correctness and .395 for rectangles (evaluated at a pre-test score of .28).

Addition Learning. To determine if there were condition differences on the addition item scores at pre-test, I ran an ANOVA on the pre-test addition scores, with test form, order, grade, and condition as fixed factors, with a full-factorial model. None of the interactions were significant so the model was re-run without them. With a main-effects model, only grade was significant (p<.0005). Since

pre-test form and order were not significant, they are not included in further analyses for the addition scores.

To determine if there were differences by condition for immediate learning of the target addition content, I ran an ANCOVA on the post-test addition scores, with condition and grade as fixed factors, pre-test addition score as a covariate, and interaction terms for condition by grade and condition by pre-test score. Neither interaction term was significant so the model was re-run without them. With the main-effects model, there was a significant effect for grade (p=.002) pre-test addition score (p<.0005), and condition (p=.046). Estimated marginal means are .273 for correctness and .220 for rectangles (evaluated at a pre-test score of .111).

To determine if there were differences by condition for retention and future learning of the target addition content, I ran an ANCOVA on the delayed-test addition scores, with condition and grade as fixed factors, post-test addition score as a covariate, and interaction terms for condition by post-test addition score and condition by grade. Neither interaction was significant so the model was re-run without them. The only significant main effect was post-test addition scores (p<.0005), with no significant effect for grade or condition (both p>.3).

To determine if there were condition differences across the whole study on the target addition content, I ran an ANCOVA on the delayed-test addition scores with condition and grade as fixed factors, pre-test addition score as a covariate, and interaction terms for condition by grade and condition by pre-test addition score. The interactions were not significant so the model was re-run without them. With main effects only, grade was significant (p=.019), as was pre-test addition score (p<.0005), with no significant effect of condition (p>.2).

Evaluation Items. In study 3, students in the grounded condition had greater pre-to-post gains than the correctness condition on the pictures-and-numbers evaluation items, while showing similar gains on the numbers-only evaluation item. To examine if this finding holds with the present study, I ran a MANOVA on the post-test scores for those two groups of items. With the pre-test scores for those items as covariates and grade and condition as fixed factors, the multivariate tests show significant effects for each pre-test score (both p<.0005), grade (p<.0005), and a marginal grade by condition interaction (p=.055). Condition is not significant (p>.8). Between-subject effects show that grade is significant for post-test numbers-only items (p<.0005) but not for pictures-and-numbers items (p>.4). The condition by grade interaction is also significant for the post-test numbers-only items (p=.02) but not for the pictures-and-numbers items (p>.8). Significant parameter estimates for post-test numbers-only score for terms including condition or grade are .114 for correctness (p=.022), -.102 for 4th grade, (p=.026) and -.183 for the interaction of correctness and 4th grade (p=.02). None of the parameter estimates for terms including condition or grade were significant for the post-test pictures-and-numbers scores. This analysis does not replicate the findings from study 3, as it finds no main effect of condition for the pictures-and-numbers evaluation items. While the interaction of grade and

condition seems to suggest that the fraction bar condition led to greater learning on the numbers-only evaluation items for 4th graders, it was not a prior hypothesis and the interaction is marginal at the multi-variate level.

Comparing the pre-to-delayed gains for the numbers-only and pictures-and-numbers evaluation items leads to similar results. This time the interaction of grade and condition was not significant at the multivariate level (p>.2) so the analysis was re-run without it. With main effects only, grade was significant (p=.001) as were the pre-test scores for each item type (p=.003 for numbers-only; p<.0005 for pictures-and-numbers). Condition was not significant (p>.8). Between-subject effects show that grade is significant for numbers-only (p<.0005) but not for pictures-and-numbers (p>.2). The parameter estimate for 4th grade for the numbers-only delayed-test score is -.142. Again, these results do not replicate the findings from study 3: condition was not significant as a main effect for the pictures-and-numbers evaluation items.

Given that there were no significant differences in pre-to-delayed learning by condition for the addition items or the pictures-and-numbers evaluation items, what accounts for the significant difference in improvement for the pre-to-delayed scores overall? One possible candidate is the pictures-only evaluation items. One analysis examined delayed-test scores excluding the addition items, pictures-only evaluation items, and pictures-and-numbers evaluation items, and another analysis examined learning on the pictures-only items. An ANCOVA on delayed-test scores without the addition items or the pictures-only or pictures-and-numbers evaluation items, with the corresponding pre-test score as a covariate and grade and condition as fixed factors shows no significant effect of condition (p>.2). The first model included main effects and interaction terms for condition by grade and condition by pre-test score, neither of which were significant (p>.4 for both). When condition by grade was removed, condition by pre-test score remained non-significant (p>.4). Condition was not significant in either of those models, or in a model that only included main effects (p>.2 for all models). In all three models, grade was marginally significant (p=.077 in the first model, p=.083 in the second, and p=.074 in the main effects model). Pre-test score was significant in all models (p<.0005).

This analysis suggests that the condition difference in pre-to-delayed gains was driven by the pictures-only evaluation items, and not other items on the assessments. Further, while there was a significant pre-test by condition interaction in the full pre-test to delayed-test analysis, without the addition items and without the two evaluation types that included pictures, that interaction is not significant. While this may have resulted from a loss of power in looking at fewer test items, it is also possible that the interaction was driven by pictures-only evaluation items. An ANCOVA on the pictures-only scores at delayed-test, with scores at pre-test as a covariate and grade and condition as fixed factors shows a significant effect of condition (p=.008), and no significant effect of either interaction term. Estimated marginal means are .650 for correctness and .776 for rectangles for the delayed-test pictures-only evaluation items, evaluated at a pre-test score of .6195. Together, these analyses indicate that the fraction bar

condition had greater gains on the pictures-only evaluation items, and this accounts for the greater overall gains from pre-test to delayed-test. This result is based on an exploratory analysis and not a prior hypothesis, and is presented here only to show that there is no evidence that the fraction bar condition demonstrated relative gains on test items that did not include fraction bars.

7.3.5 Discussion

Process measures indicate that students had different experiences with each of the three tutor conditions. There were significant differences between all three conditions for the number of regular tutor problems solved and for the time taken per regular problem. The correctness tutor was the easiest to work with on these measures (most problems solved with the least amount of time per problem), followed by grounded plus correctness and then virtual manipulatives. These results make sense given the design of the three conditions: grounded plus correctness and virtual manipulatives had less time overall for the regular tutor problems because they were given the fraction bar pre-instruction. Further, both tutors with fraction bars allowed students to go down incorrect paths, such as converting fractions unnecissarily and adding when the two fractions did not have the same denominators. For the number of hints requested per problem, the means for each condition follow the same pattern (correctness requsted the fewest hints per problem, followed by grounded plus correctness and then virtual manipulatives), but the only significant differences are between virtual manipulatives and the other two conditions. Unsurprisingly, this indicates that the fraction bars alone are not as easy for students to interpret as the correctness feedback.

While the three conditions led to large differences in learning expreiences with the tutors, they did not lead to large differences in learning outcomes. In analyses of the assesment scores for all three conditions, there were no significant differences in immediate learning (pre to post). There was a marginal difference between correctness and virtual manipulatives for retention and future learning (post to delayed) in favor of virtual manipulatives. For learning across the whole study (pre to delayed), while there were significant differences in learning by condition, non of the pairwise comparisons between conditions were significant. However, there was a significant condition by pre-test score interaction, suggesting that students with higher pre-test scores would benefit most from the correctness condition. For addition learning, there was a significant difference between correctness and virtual manipulatives, in favor of correctness, for immediate learning. However, there were no significant differences in addition learning by condition from post- to delayed-test or over the entire study (pre to delayed). When the grounded plus correctness and virtual manipulatives conditions are collapsed into one Fraction Bar condition, there are no significant differences in immediate learning, but there are significant differences in post-to-delayed learning and in learning across the entire study. Both of these differences show benefits for the fraction bar condition. However, for learning across the

entire study, there was a significant condition by pre-test interaction, suggesting that that students with higher pre-test scores would benefit most from the correctness condition. For the target addition content, correctness showed a benefit over fraction bars for immediate learning. However, there were no significant differences by condition for addition learning over the whole study. On the evaluation items, 4th graders showed greater immeditae learning with the fraction bar condition, while 5th graders showed greater immediate learning with the Correctness condition. There were no significant effects for condition on the evaluation items from pre-test to delayed-test.

While the original hypotheses predicted differences in learning outcomes between correctness and each of the other two conditions, there was no evidence that correctness differed from grounded plus correctness. While the differences in process measures show that the two tutor designs had an effect on students' actions, that may not have translated to differences in students' thinking.

The results for post-to-delayed learning show a marginal difference between virtual manipulatives and correctness, suggesting that the greater difficulties encountered by the virtual manipulatives students may have led to more robust learning. The interface design of the virtual manipulatives tutor was hypothesized to draw students' attention away from the fraction symbols, and therefore encourage the learning of procedures students that worked easily on fraction bars but would not transfer redily to fraction symbols. However, students still engaged with the fraction symbols in the virtual manipulatives tutor. While students could not type into the converted or sum fraction input areas once they had manipulated the fraction bars, students could type in numbers beforehand. Even though these numbers would be overwritten when input was provided to the fraction bars, the vast majority of the students in the virtual manipulatives condition inputted numbers at least once. This behavior suggests that the interface design did not draw students' focus away from the numeric sumbols. While the interface was intended to have the symbolic numbers shown as feedback when students manipulated the fraction bars, students may have perceived the buttons to be the input device for both the fraction symbols and the fraction bars. If the input mode does not change students' allocation of attention between the two representations, there should be no difference in learning between grounded feedback and virtual manipulatiaves. However, one input mode may be more cumbersome than the other, which may lead to differences in learning by slowing students' progression through the tutor.

7.4   Conclusion

In comparing correctness feedback to grounded plus correctness and virtual manipulatives, Study 5 found hints of more robust learning with the two tutors that included fraction bars, especially so for students with low prior knowledge. For evaluating the importance of input mode to grounded feedback, this study suggests that input mode may only be important to the extent that it affects

students' allocation of attention between the input and feedback representations. Redesigning the virtual manipulatives tutor so that the buttons do not show numbers and so the students cannot enter numbers in the converted and sum fraction areas at any point may be a better test of input mode. Those aspects of the tutor design may have drawn students' attention to the symbolic representation, when the purpose of the study was to examine students' learning when their attention was focused on the more concrete representation.

8   Conclusion

This thesis presents grounded feedback, a use of multiple representations whereby the student's inputs, in the target, to-be-learned representation, are reflected in a more concrete representation that is easier for the student to reason with. Chapter 1 defines grounded feedback with four features:

1) The feedback is intrinsic to the domain and reflects the students' inputs.

2) Students can easily envision the feedback state that indicates a correct answer to a given problem. Therefore, by examining the feedback, students can evaluate for themselves if their answers are correct or not.

3) Students do not directly manipulate the feedback representation. Instead, the inputs are in a format that matches the domain learning goals.

4) The feedback affords meaningful inferences on errors, beyond the indication that an action was incorrect. By examining the feedback representation and its correspondence to the input representation, the student can extract information about the nature of the error.

The concrete representation was hypothesized to help students evaluate the outcome of their procedures in light of the concepts that underlie them. By facilitating this evaluation step instead of performing it for the student, grounded feedback is hypothesized to support students in connecting concepts and procedures and therefore promote robust learning. Having students' inputs be in the abstract representation instead of the concrete representation was hypothesized to promote transfer to abstract-only contexts. Finally, beyond

supporting students in evaluating the correctness of an action, grounded feedback is hypothesized to facilitate students' inferences about the domain (e.g., that adding the numerators and denominators of two positive addends yields a sum that is incorrect because it is too small). While grounded feedback has been implemented in several educational technology systems, it has not previously been explicitly defined.

Is grounded feedback effective for learning? Studies 2 and 3, along with prior experiments (Mathan & Koedinger, 2005; Nathan, 1998), indicate that it is. When compared to immediate step-level correctness feedback, grounded feedback led to greater learning from pre-test to delayed-test (in Study 2) and greater improvement on transfer tasks from pre-test to post-test (in Study 3, which did not have a delayed-test). These greater gains, in both studies, did not come at the expense of procedural skill. Even though students in the grounded conditions practiced fraction addition with two representations, the demonstrated improvement with symbols-only fraction addition was not significantly different from the symbols-only correctness condition (both from pre-test to post-test and from pre-test to delayed-test in Study 2, and from pre-test to post-test in Study 3).

8.1 Is Each Feature of Grounded Feedback Important?

The studies in this thesis have partially evaluated the first three features of grounded feedback, and a comparison between these findings and those of prior work in other domains offers a hypothesis about the last feature.

The first feature, providing feedback with an intrinsic representation, is supported by studies 2, 3, and 5, which found benefits for the conditions that included fraction bars over the symbols-only control condition. Even when students struggled in using feedback to evaluate their own work (as in Study 2), they still benefitted from it. The difficulty factor assessments in Chapter 4 shows how to evaluate if a representation is intrinsic and appropriate for grounded feedback: determining that students can perform a domain-relevant task with the feedback representation alone (e.g., deciding if an equation is true or not with the pictures-only representation), and that the task is made easier with the feedback representation as compared with the symbolic representation alone (e.g., the evaluation task was easier in the pictures-and-numbers format compared with numbers-only).

The motivating hypothesis behind the second feature, that students can use the grounded representation to evaluate their own work, was that self-evaluation would help students connect their prior conceptual knowledge to the task at hand. Specifically, the necessity of self-evaluation, prompted by the absence of step-level correctness feedback, would help students actively engage their own knowledge. This mechanism was not born out by Study 4, which found that students benefitted equally from grounded feedback, whether step-level correctness feedback was included or not. Additionally, pairwise analyses for the

three conditions in Study 5 did not indicate differences in learning between the virtual manipulatives condition (which did not include step-level correctness feedback) and the grounded plus correctness condition (which did). Further research is necessary to determine if including step-level correctness feedback influences how students engage with the grounded feedback.

The motivating hypothesis behind the third feature was that if students could manipulate the concrete representation directly, they would have no motivation to engage with the harder-to-interpret target representation. This would harm students' learning since procedures on concrete representations often do not transfer easily to procedures on abstract symbols (Resnick & Omanson, 1987; Sarama & Clements, 2009; Uttal et al., 2013). The results from Study 5 do not support the first part of this hypothesis: even though students in the virtual manipulatives condition manipulated the rectangles, they still engaged with the symbolic numbers. Evidence for this engagement is found in the log data.

A bug in the interface design allowed students to input numbers for the converted and sum fractions when those interfaces first appeared. Those inputs would be overwritten once the student began to set the value for the fraction bar (which was necessary for advancing to the next problem). Students not only engaged with the symbolic representation, they did so even when it meant performing additional, unnecessary steps. Some of this engagement may have been promoted by details of the interface design: the input buttons for the fraction bars were labeled with numeric symbols.

Additionally, students may have attended to the symbolic representation because of the particular demands of the fraction addition task. While the concrete representation makes *evaluating* the equations easier, it is not trivial to *solve* such equations with the concrete representation alone. In particular, when dealing with large denominators (e.g., $7/24 + 1/18$), it is easier to confirm that two fractions have been converted to the same denominator (e.g., $21/72 + 4/72$) when looking at the symbolic representation rather than the concrete one. While prior work has found poor transfer between learning with a concrete representation and performance with an abstract one (and visa versa; Uttal et al., 2013), I am not aware of work that examined the importance of input modality when both representations are presented together. The results from Study 5 suggest that students will benefit when they attend to both representations, and there are nuances in designing instruction that manipulates students' attention. In retrospect, attention on the abstract representation could have been reduced if the input buttons were not labeled with numbers and if the bug that afforded numeric input was eliminated. Further, it is likely that features of the domain will determine how students allocate their attention. For example, if the task had been finding equivalent fractions instead of fraction addition, students may not have engaged with the symbolic representation. Future work should continue to investigate how students allocate their attention between representations when concrete and abstract representations are both available.

The last feature of grounded feedback, that students can make meaningful inferences about the domain by coordinating between the input and feedback

representations, was not evaluated experimentally in this thesis. However, a comparison of the results from studies with the fraction addition tutors and the Intelligent Novice Spreadsheet tutor (Mathan & Koedinger, 2005) and ANIMATE (Nathan, 1998), suggests that better coordination of the input and feedback representations leads to better learning. While grounded feedback in the fraction addition tutors led to improved long-term learning (pre-to-delayed-test) and improved learning on transfer tasks compared to a correctness-feedback control, none of the studies in this thesis found benefits for grounded feedback for immediate learning of symbolic fraction addition. In contrast, Mathan and Koedinger (2005) found immediate benefits for the intelligent novice spreadsheet tutor, over the immediate correctness-feedback control, not only for transfer tasks but also for the target task. Nathan (1998) also found immediate relative benefits for the target algebra-symbolization task, although those benefits were only shown for two of the three problem types.

An ordering of the three research projects by the robustness of their findings (spreadsheets, algebraic symbolization, and then fraction addition) appears to correspond with the level of coordination students achieved in each setting, whether with support or own their own. The intelligent novice spreadsheet tutor provided explicit support for mapping between the formula input, the cells that it referenced, and the resulting value (Mathan & Koedinger, 2005). This coordination support was provided as part of the intelligent novice guidance: if students made an error and could not identify or correct it on their own, the tutor would provide the coordination support to help students see the error not just in terms of the feedback representation, but also in terms of how the error manifested in the input representation. Only after showing the mapping between the cause (the incorrect cell reference) and the effect (the incorrect cell value) did the tutor provide guidance in correcting the error. Therefore, the intelligent novice tutor can be interpreted as not just offering a context for intelligent novice errors to occur and to be corrected by the student, but also as offering explicit instruction and practice with coordination.

The ANIMATE tutor did not provide explicit support for interpreting the animation feedback, but students often coordinated the representations on their own as demonstrated both by a reduction of error rates between pre-test problems and tutor problems and by a large number of self-corrections during tutoring in response to the animations (Nathan, 1991). However, students seemed better able to coordinate the representations when the problems involved travel or interest rates rather than work (Nathan, 1998). The crucial difference appears to be the difference between the values presented in the problem and the values needed in the equations. With travel problems, students were given the vehicles' speed in miles per hour, which could be directly inputted to an equation such as *miles traveled = speed * hours traveled.* However, with work problems, students were given a productivity level in hours needed to complete a task (e.g., "Tom can paint the fence in two hours"; Nathan, 1991, p.102), while the equations in the tutor interface required the amount of work that could be completed in one hour (e.g., *work completed = work performed per hour * hours worked*).

Therefore, the work problems required students to use the reciprocal of the given value (.5 fences per hour instead of 2 hours per fence), while the travel and investment problems allowed students to use the rates as they were given in the problem statement. The necessity of using the reciprocal made it more difficult for students to coordinate the symbolic and animated representations: when students used the given value they could tell that the animation showed the wrong rate, but the animations did not provide enough support for the student to know what to try next (Nathan, 1998). Therefore, for the work problems, the animation feedback offered no more information than correctness feedback, which provides an explanation for why grounded feedback led to better learning than correctness feedback for the travel and investment problems but not the work problems.
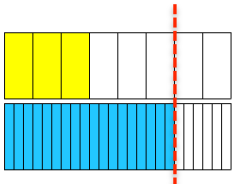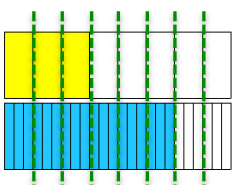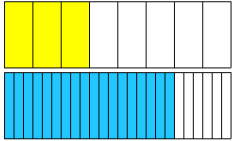
Of the three domains, students had the most difficulty coordination the representations in the fraction addition tutor. The difficulty factor assessments discussed in Chapter 4 show that proficiency in evaluating a fraction addition equation is *reduced* when symbols are included with the fraction rectangles. The analogous situation in the other domains would be for students to be *worse* at evaluating the correctness of a spreadsheet formula if they saw the formula along with the calculated values for each cell (as opposed to just seeing the calculated values), or for a student to be *worse* at evaluating if an animation matched a story problem if they saw the algebraic equations along with the animation (as opposed to the animation alone). Although Studies 2 and 3 show benefits for grounded feedback in the domain of fraction addition, it is likely that grounded feedback will be more effective when students are better able to coordinate between the two representations.

An issue related to the coordination of representations is the students' prior conceptual knowledge. The second difficulty assessment in Chapter 4 suggests that one reason students may have trouble coordinating the symbolic and concrete fraction representations is that they do not have a solid foundation of qualitative inference rules for addition. Specifically, students did not find it obvious that the sum of two positive fractions is larger than each addend alone. While this concept is not necessary, in the strictest sense, for coordinating the representations, it does seem to be relevant. Neither of the prior experiments (Mathan & Koedinger, 2005; Nathan, 1998) report gaps in students' prior conceptual knowledge that would impinge on their ability to interpret the feedback representation.

8.2 An Ideal Model of Coordination with Grounded Feedback

Practice with the grounded feedback tutor transfers to symbolic addition as evidenced by students' improvement on the fraction addition items over the course of each study. What cognitive processes might account for this transfer? Table 8.1 illustrates potential cognitive steps that students might take after

making an error when using grounded feedback, with the fraction conversion from Chapter 5 as an example.

Domain-General Self-Monitoring	Fraction Conversion Example	Grounded Feedback
1. Detect a discrepancy between the actual state and the expected state, in terms of both representations.	Expected state: the converted fraction will be equivalent to the addend, so both fraction bars will have the same amount filled in Actual state: 18/24 does not line up with 3/8, meaning that 18/24 does not equal 3/8	
2. For all of the other ways that the answer can vary, identify if any are correct, in terms of both representations	The dividing lines for the addend line up with those of the converted fraction, meaning that the addend can be converted to 24ths. This is the correct denominator because it works for both 3/8 and 1/3 (the other addend).	
3. Identify which aspects of the answer should be changed, for both representations	The dividing lines are correct but the total amount colored in is wrong, meaning that only the numerator should be changed	
4. Identify how those aspects should be changed	Too much is colored in; numerator should be smaller	
5. Execute the change, repeating until the original discrepancy is resolved	Input smaller numerators until the magnitudes of the two fractions are equal	
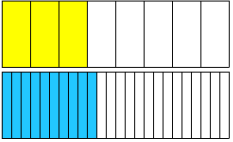
Table 8.1: Ideal domain-general self-monitoring execution with fractions and supported by grounded feedback.

The domain-general cognitive steps in table 8.1 express error detection and correction at a high level. The corresponding problem-specific steps for fraction conversion illustrate how a student could use the grounded feedback to carry out those domain-general steps without prior knowledge of how to convert symbolic fractions procedurally. These cognitive moves illustrate how grounded feedback can draw on conceptual knowledge for evaluation and reinforce the connection between the conceptual knowledge and procedural steps. For the first step, identifying a discrepancy, the student must draw on conceptual knowledge that equivalent fractions have the same magnitude. This may reinforce the connection between conceptual knowledge and the larger fraction addition procedure: instead of adding the original addends directly, one converts and adds the converted fractions, which is reasonable only because the converted fractions have the same magnitude as the original addends. For the second step, identifying correct aspects of the current answer, the student again draws on conceptual knowledge: to find a denominator for both converted fractions, it must be divisible by the denominator for each original addend. In terms of the grounded feedback, this divisibility relationship is represented by the alignment between the dividing lines for the original addends and the dividing lines for the converted fraction. For the third step, identifying which aspect of the current answer should be changed, the student draws on conceptual knowledge regarding the role of the numerator and denominator in determining the magnitude of the fraction: the denominator determines the size of the fraction pieces, while the numerator determines how many are present. The execution steps provide a check for the student's reasoning: making the numerator smaller should yield a new fraction that is smaller, but should not change the size of the pieces. Reinforcing the connection between the concepts and procedures may give the student more confidence that the procedural steps are correct, and may also help the student re-construct the procedure if it was not memorized completely.

For all studies that included the correctness feedback tutor (studies 2, 3, and 5), students in the correctness condition solved more tutor problems than students in the conditions that included fraction bars. Students in the correctness conditions therefore had more opportunities to practice the fraction addition procedure with different addends, and this procedural practice likely caused the fraction addition learning. Students in the grounded conditions also improved, but through a different path. They solved fewer problems, and therefore did not have as much procedural practice. (While they spent more time per problem, some of that time was spend on conceptual tasks and thus they had less total procedural practice time.) Instead, their learning likely stemmed from gains in both conceptual and procedural knowledge, and practice connecting the two. This combined practice produced a kind of less-is-more effect on procedural outcomes whereby they learned just as much with fewer total practice problems and less time allocated purely toward procedural practice.

8.3   Limitations

While this thesis demonstrates benefits for learning with grounded feedback, it does not provide evidence for a complete mechanism for *how* students learn with grounded feedback. While one hypothesized process for connecting concepts and procedures is presented above (section 8.2), the current studies cannot provide evidence that students engage in such processes, either implicitly or explicitly. One hypothesis is that grounded feedback will be more effective for students who are better able to coordinate between the input and feedback representations. One assessment of this coordination ability was the set of evaluation items that included both fraction bars and symbols (from the Difficulty Factor Assessments in Chapter 4 and in the pre- and post-tests for Studies 3, 4, and 5). If these items test knowledge that is important for learning from grounded feedback but not from correctness feedback, then performance on these items should predict learning for students working with grounded feedback, but not for students working with correctness feedback.

To test this assumption, I ran an ANCOVA on the post-test scores for Study 3, with the full pre-test score as a covariate, and the score on the pictures-and-numbers evaluation items as an additional covariate. Grade and condition were included as fixed factors, and a condition by evaluation-item-pre-test-score interaction. The interaction was not significant ($p>.4$), indicating that performance on the pictures-and-numbers evaluation items did not differentially predict learning between the two conditions. Another analysis examined post-test addition scores, with pre-test addition scores and pre-test pictures-and-numbers evaluation scores as covariates, grade and condition as fixed factors, and the condition by evaluation-item-pre-test-score interaction. Again, the interaction was not significant ($p>.9$), indicating that the pictures-and-numbers evaluation items did not differentially predict addition learning between the two conditions. Similar analyses for Study 5 also did not reveal a significant interaction.

One interpretation of this negative result is that coordination may not be the key to learning from grounded feedback. Alternatively, the test items may have been overly coarse for measuring students' coordination. Additionally, students may respond differently to grounded feedback during problem solving than they do to a grounded representation on an assessment. Future work should probe the specific mechanisms for how students learn from grounded feedback, and what types of assessment items will accurately predict such learning.

Finally, students in conditions with fraction bars requested more hints than students in the correctness conditions (studies 2, 3, and 5), and it is possible that greater engagement with the hints caused students' learning, either by supporting students' interpretation of the grounded feedback or independently. An *in vivo* study with a tutor that did not provide immediate correctness feedback or grounded feedback was thought to be unethical because it was hypothesized to lead to unproductive floundering. A lab study with such a tutor may help distinguish between learning benefits from the hints alone and learning benefits from hints plus grounded feedback.

8.4   Contributions and Future Work

This thesis defined and began to evaluate grounded Feedback, a use of dual representations that was present in prior work but had not been explicitly defined or fully evaluated. This thesis provides evidence that grounded feedback is effective for learning. However, this thesis also shows that grounded feedback is not uniformly effective across domains, and indicates that students' prior conceptual knowledge and their ability to coordinate the input and feedback representations may influence grounded feedback's effectiveness. Future work should continue to explore how features of the domain and students' prior knowledge affect the benefits of grounded feedback. Further, future work should examine if explicit support for coordination and metacognitive support for invoking a "self-critic" may enhance students' learning with grounded feedback. As discussed in Chapter 4, a self-critic uses conceptual knowledge to evaluate the outcome of a procedure (e.g., a self-critic may checks the result of an arithmetic procedure using the qualitative inference rules that the sum of two positive addends must be larger than either alone). While domain-specific support for coordinating the input and feedback representations will likely help the self-critic function, students may also benefit from metacognitive support for invoking it in the first place. This domain-general support could prompt students to check their work, to think about what conceptual knowledge they can use to check their work, and to consider what information present in the tutoring interface has relevance for those concepts.

The domain-general self-monitoring steps in table 8.1 are metacognitive steps that form a procedure for self-monitoring(Zimmerman & Campillo, 2003). They are metacognitive in the sense that they reflect on responses produced by cognition – in this case, coordinating between cognition that results from

considering both representations. As procedural steps, students will likely learn them better with practice and feedback. These steps provide a sketch of a metacognitive model that could provide the basis for metacognitive model tracing and tutoring (Roll, Aleven, McLaren, & Koedinger, 2011; Walker, Koedinger, McLaren, & Rummel, 2006). Further, while this thesis examined grounded feedback when students interacted with it individually, students may benefit from working collaboratively, discussing the feedback with a partner and justifying their self-evaluations (cf. Walker, Rummel, & Koedinger, 2014).

This thesis also investigated students' evaluation of fraction addition equations and provides evidence that middle school students may lack foundational concepts for fraction addition. In particular, while students seem to understand addition with fraction bars alone, the presence of fraction symbols reduces students' performance. Further, it was not obvious to 5th graders that the sum of two positive fractions is larger than either of its addends alone. These results indicate gaps in students' knowledge of the relationship between fraction symbols and the magnitudes they refer to, as well as the role of magnitude in addition. These findings suggest that middle school students may need more instruction on these concepts.

References

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, *33*(2-3), 131–152. doi:10.1016/S0360-1315(99)00029-9

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the Effects of Different Multiple Representational Systems in Learning Primary Mathematics. *Journal of the Learning Sciences*, *11*(1), 25–61. doi:10.1207/S15327809JLS1101_2

Aleven, V., Mclaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A New Paradigm for Intelligent Tutoring Systems: Example-Tracing Tutors. *International Journal of Artificial Intelligence in Education*, *19*(2), 105–154. Retrieved from http://iospress.metapress.com/index/X3760U67H22LM111.pdf

Beckmann, S. (2004). Solving Algebra and Other Story Problems with Simple Diagrams: a Method Demonstrated in Grade 4–6 Texts Used in Singapore. *The Mathematics Educator*, *14*(1), 42–46.

Bjork, E. L., & Bjork, R. (2009). Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (pp. 55–64). New York, NY: Worth Publishers.

Bodemer, D., Plötzner, R., Bruchmüller, K., & Häcker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. *Instructional Science*, *33*, 73–95.

Booth, J. L., & Koedinger, K. R. (2011). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. *British Journal of Educational Psychology*, *82*(3), 492–511. doi:10.1111/j.2044-8279.2011.02041.x

Byrnes, J. P., & Wasik, B. A. (1991). Role of Conceptual Knowledge in Mathematical Procedural Learning. *Developmental Psychology*, *27*(5), 777–786.

Clark, R. C., & Mayer, R. E. (2011). *E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning* (3rd ed.). San Francisco, CA: Pfeiffer.

Dugdale, S. (1992). The Design of Computer-Based Mathematics Instruction. In J. H. Larkin & R. W. Chabay (Eds.), *Computer-Assisted Instruction and Intelligent Tutoring Systems* (pp. 11–45). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Fyfe, E. R., McNeil, N., Son, J., & Goldstone, R. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, *26*(1), 9–25.

Gomoll, K. (1990). Some Techniques for Observing Users.pdf. In B. Laurel (Ed.), *The art of human-computer interaface design* (pp. 85–90). Reading, MA: Addison-Wesley.

Heffernan, N. T., & Koedinger, K. R. (1998). A Developmental Model for Algebra Symbolization: The Results of a Difficulty Factors Assessment. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 484–489). Hillsdale, New Jersey: Erlbaum.

Horwitz, P., & Barowy, B. (1994). Designing and Using Open-Ended Software to Promote Conceptual Change. *Journal of Science Education and Technology*, *3*(3), 161–185. doi:10.1007/BF01575178

Izsak, A. (2000). Inscribing the Winch: Mechanisms by Which Students Develop Knowledge Structures for Representing the Physical World With Algebra. *Journal of the Learning Sciences*, *9*(1), 37–41.

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science*, *38*(6), 523–550. doi:10.1007/s11251-009-9093-x

Koedinger, K. R., & Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review*, *19*(3), 239–264. doi:10.1007/s10648-007-9049-0

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: evidence from algebra problem solving. *Cognitive Science*, *32*(2), 366–97. doi:10.1080/03640210701863933

Lave, J. (1988). *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life.* Cambridge, UK: Cambridge University Press.

Lovett, M. (1998). Cognitive task analysis in service of intelligent tutoring system design: A case study in statistics. In *Intelligent Tutoring Systems* (pp. 234–243). Springer Berlin Heidelberg.

Mathan, S., & Koedinger, K. R. (2003). Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In H. U. Hoppe, M. F. Verdejo, & J. Kay (Eds.), *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies, Proceedings of AI-ED 2003* (pp. 13–18). Amsterdam, the Netherlands: IOS Press.
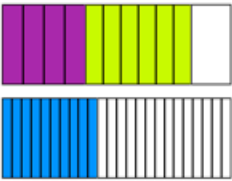
Mathan, S., & Koedinger, K. R. (2005). Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. *Educational Psychologist*, *40*(4), 257–265. doi:10.1207/s15326985ep4004_7

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-School Students' Understanding of the Equal Sign: The Books They Read Can't Help. *Cognition and Instruction*, *24*(3), 367–385. doi:10.1207/s1532690xci2403_3

Nathan, M. J. (1991). A simple learning environment improves mathematical reasoning. *Intelligent Tutoring Media*, *2*(3-4), 101–111.

Nathan, M. J. (1998). Knowledge and Situational Feedback in a Learning Environment for Algebra Story Problem Solving. *Interactive Learning Environments*, *5*(1), 135–159.

Ohlsson, S. (1996). Learning from Performance Errors. *Psychological Review*, *103*(2), 241–262.

Padalkar, S., & Hegarty, M. (2012). Improving Representational Competence in Chemistry with Model-Based Feedback. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 2162–2167). Austin, TX: Cognitive Science Society.

Powers, W. T. (1973). *Behavior: The Control of Perception*. New York, NY: Hawthorne.

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense Making Alone Doesn't Do It: Fluency Matters Too! ITS Support for Robust Learning with Multiple Representations. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 174–184). Springer Berlin Heidelberg. doi:10.1007/978-3-642-30950-2

Resnick, L. B., & Omanson, S. F. (1987). Learning to Understand Arithmetic. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 41–95). Hillsdale, New Jersey: Erlbaum.

Rittle-Johnson, B., & Koedinger, K. R. (2001). Using cognitive models to guide instructional design: The case of fraction division. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 857–862). Mahwah, NJ: Erlbaum.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*(2), 267–280. doi:10.1016/j.learninstruc.2010.07.004

Roschelle, J., Kaput, J., & Stroup, W. (2000). SimCalc: Accelerating Students' Engagement with the Mathematics of Change. In M. J. Jacombson & R. B. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs for technologies of learning* (pp. 47–75). Hillsdale, New Jersey: Earlbaum.

Sarama, J., & Clements, D. H. (2009). "Concrete" Computer Manipulatives in Mathematics Education. *Child Development Perspectives*, *3*(3), 145–150. doi:10.1111/j.1750-8606.2009.00095.x

Schoenfeld, A. H. (1988). When Good Teaching Leads to Bad Results: The Disasters of "Well Taught" Mathematics Courses. *Educational Psychologist*, *23*(2), 1–22.

Schwartz, D., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, *22*(2), 129–184. doi:10.1207/s1532690xci2202_1

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. doi:10.3102/0034654307313795

Stampfer, E., & Koedinger, K. R. (2012). Tradeoffs between Immediate and Future Learning. *Paper presented at the European Association for Learning and Instruction.* Bari, Italy.

Stampfer, E., & Koedinger, K. R. (2013). When seeing isn't believing: Influences of prior conceptions and misconceptions. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1384–1389). Berlin, Germany: Cognitive Science Society.

Suh, J., Moyer, P. S., & Heo, H.-J. (2005). Examining Technology Uses in the Classroom: Developing Fraction Sense Using Virtual Manipulative Concept Tutorials. *Journal of Interactive Online Learning*, *3*(4), 1–21.

Uttal, D. H., Amaya, M., Maita, M. del R., Hand, L. L., Cohen, C. A., O'Doherty, K., & Deloache, J. S. (2013). It works both ways: Transfer difficulties between manipulatives and written subtraction solutions. *Child Development Research*, *2013*. doi:10.1155/2013/216367

Walker, E., Koedinger, K., McLaren, B., & Rummel, N. (2006). Cognitive tutors as research platforms: Extending an established tutoring system for collaborative and metacognitive experimentation. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*. Springer Berlin Heidelberg. doi:10.1007/11774303_21

Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *International Journal of Artificial Intelligence in Education*, *24*(1), 33–61.

Wood, H., & Wood, D. (1999). Help seeking, learning and contingent tutoring. *Computers & Education*, *33*(2-3), 153–169. doi:10.1016/S0360-1315(99)00030-5

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., … Ogbuehi, P. (2012). *Improving Mathematical Problem Solving in Grades 4 Through 8*.

Yerushalmy, M. (1991). Effects of Computerized Feedback on Performing and Debugging Algebraic Transformations. *Journal of Educational Computing Research*, *7*(3), 309–330.

Zimmerman, B. J., & Campillo, M. (2003). Motivating Self-Regulated Problem Solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The Psychology of Problem Solving* (pp. 233–262). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511615771

Appendix A: Test form A

This test form was used in studies 3, 4, and 5, with students randomly assigned to see the questions in forward or reversed order.

$$\frac{4}{11} + \frac{6}{13}$$

Is this correct?

$$\frac{4}{11} + \frac{6}{13} = \frac{10}{23}$$

$$\frac{10}{23}$$

Choose the answer that goes in the blank:

$\frac{10}{23}$ is ____ .

○ too small
○ correct
○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{2}{12} + \frac{5}{12} = \frac{}{}$$

✔ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{5}{13} + \frac{4}{13} = \frac{\boxed{}}{\boxed{}}$$ ✔ Done

You may use this area for scratch work:

Is this correct?

$$\frac{2}{5} + \frac{3}{9} = \frac{5}{45}$$

Choose the answer that goes in the blank:

$$\frac{5}{45}$$ is _____ .

○ too small
○ correct
○ too big   ✔ Done

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{5}{9} + \frac{3}{10} = \frac{\boxed{}}{\boxed{}}$$ ✔ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{3}{21} + \frac{1}{3} = \frac{\boxed{}}{\boxed{}}$$

✔ Done

You may use this area for scratch work:

Fill in the numerator to make the fractions equivalent.

$$\frac{3}{6} = \frac{\boxed{}}{18}$$

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{4}{6} + \frac{1}{4} = \frac{\boxed{}}{\boxed{}}$$

✔ Done

You may use this area for scratch work:

Is this correct?

$$\frac{2}{11} + \frac{1}{2} = \frac{3}{13}$$

Choose the answer that goes in the blank:

$\frac{3}{13}$ is _____ .

- ○ too small
- ○ correct
- ○ too big

✔
Done

This addition is true. The questions below ask about the numbers in the equation.

$$\frac{6}{38} + \frac{8}{19} = \frac{33}{57}$$

Is $\frac{33}{57}$ greater than $\frac{6}{38}$?	Is $\frac{33}{57}$ greater than $\frac{8}{19}$?
○ Yes	○ Yes
○ No	○ No
○ Not enough information	○ Not enough information

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{2}{3} + \frac{1}{5} = \frac{}{}$$

✔
Done

You may use this area for scratch work:

139

A bag of jellybeans has 10 jellybeans in it.

Michelle has $\dfrac{3}{10}$ of a bag.   Barack has $\dfrac{4}{10}$ of a bag.

What fraction of a bag of jellybeans do they have all together?

$\dfrac{\boxed{}}{\boxed{}}$

Done
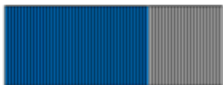
$\dfrac{1}{4} + \dfrac{7}{17}$

$\dfrac{48}{65}$

Is this correct?

$$\dfrac{1}{4} + \dfrac{7}{17} = \dfrac{48}{65}$$

Choose the answer that goes in the blank:

$\dfrac{48}{65}$ is _____ .

○ too small
○ correct
○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\dfrac{1}{2} + \dfrac{4}{11} = \dfrac{\boxed{}}{\boxed{}}$$
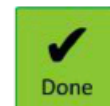
✔ Done

You may use this area for scratch work:

140

Is this correct?

$+$ $=$

Choose the answer that goes in the blank:

is _____ .

○ too small
○ correct
○ too big

✔ Done

Do NOT add these fractions. > means "is greater than".

Is this statement true or false?

$$\frac{16}{51} + \frac{33}{47} > \frac{33}{47}$$

○ True
○ False

✔ Done

Is this correct?

$+$ $=$

Choose the answer that goes in the blank:

is _____ .

○ too small
○ correct
○ too big

✔ Done

Fill in the numerator to make the
fractions equivalent.

$$\frac{4}{5} = \frac{\boxed{}}{15}$$

✔
Done
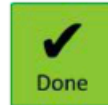
Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{2}{5} + \frac{11}{20} = \frac{\boxed{}}{\boxed{}}$$

✔
Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{1}{4} + \frac{3}{16} = \frac{\boxed{}}{\boxed{}}$$

✔
Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{3}{4} \ + \ \frac{1}{7} \ = \ \frac{\ \ \ \ \ }{\ \ \ \ \ }$$ ✔ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{5}{13} \ + \ \frac{1}{3} \ = \ \frac{\ \ \ \ \ }{\ \ \ \ \ }$$ ✔ Done
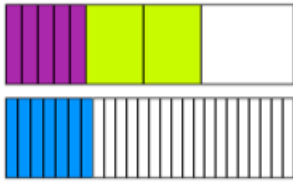
You may use this area for scratch work:

Is this correct?

Choose the answer that goes in the blank:

 is _____ .

- ○ too small
- ○ correct
- ○ too big

Done

Is this correct?

$$\frac{4}{6} + \frac{2}{9} = \frac{16}{18}$$

Choose the answer that goes in the blank:

$\frac{16}{18}$ is _____ .

- ○ too small
- ○ correct
- ○ too big

Done

This addition is true. The questions below ask about the numbers in the equation.

$$\frac{15}{22} + \frac{6}{33} = \frac{57}{66}$$

Is $\frac{57}{66}$ greater than $\frac{15}{22}$?

○ Yes
○ No
○ Not enough information

Is $\frac{57}{66}$ greater than $\frac{6}{33}$?

○ Yes
○ No
○ Not enough information

✔
Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{3}{11} + \frac{7}{11} = \frac{\boxed{}}{\boxed{}}$$

✔
Done

You may use this area for scratch work:

$$\frac{2}{3} + \frac{3}{25}$$
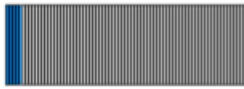
$$\frac{5}{75}$$

Is this correct?

$$\frac{2}{3} \quad + \quad \frac{3}{25} \quad = \quad \frac{5}{75}$$

Choose the answer that goes in the blank:

$$\frac{5}{75}$$ is _____ .

○ too small
○ correct
○ too big

✔ Done

Do NOT add these fractions. > means "is greater than".

Is this statement true or false?

$$\frac{41}{66} \quad + \quad \frac{19}{35} \quad > \quad \frac{41}{66}$$

○ True
○ False

✔ Done

There are 9 blocks. Color in $\frac{1}{3}$ of them.
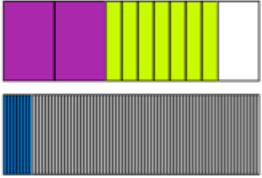
To color a block, click on it.
Click on it again to uncolor it.
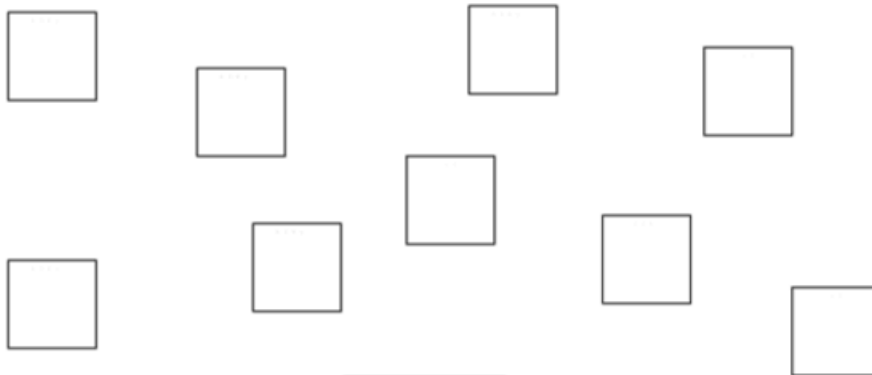
Done

Appendix B: Test form B

This test form was used in studies 3, 4, and 5, with students randomly assigned to see the questions in forward or reversed order.

$$\frac{2}{5} + \frac{7}{16}$$

$$\frac{9}{80}$$

Is this correct?

$$\frac{2}{5} + \frac{7}{16} = \frac{9}{80}$$

Choose the answer that goes in the blank:

$\frac{9}{80}$ is ____ .

- ○ too small
- ○ correct
- ○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{6}{9} + \frac{4}{27} = \frac{}{}$$

✔ Done

You may use this area for scratch work:

$$\frac{4}{13} + \frac{1}{5}$$
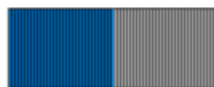
Is this correct?

$$\frac{33}{65}$$

$$\frac{4}{13} + \frac{1}{5} = \frac{33}{65}$$

Choose the answer that goes in the blank:

$\frac{33}{65}$ is _____ .

○ too small

○ correct

○ too big

✔ Done

Is this correct?

Choose the answer that goes in the blank:

is _____ .

○ too small

○ correct

○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{5}{16} \ + \ \frac{6}{16} \ = \ \frac{}{}$$

✔ Done

You may use this area for scratch work:

$$\frac{3}{22} + \frac{3}{4}$$

$$\frac{3}{26}$$

Is this correct?

$$\frac{3}{22} \ + \ \frac{3}{4} \ = \ \frac{6}{26}$$

Choose the answer that goes in the blank:

$$\frac{6}{26}$$ is _____ .

○ too small
○ correct
○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{3}{14} + \frac{2}{14} = \frac{\boxed{}}{\boxed{}}$$ ✔ Done

You may use this area for scratch work:

Fill in the numerator to make the fractions equivalent.

$$\frac{2}{3} = \frac{\boxed{}}{6}$$ ✔ Done

Fill in the numerator to make the fractions equivalent.

$$\frac{1}{4} = \frac{\boxed{}}{12}$$ ✔ Done

Is this correct?

$$\frac{2}{9} + \frac{1}{7} = \frac{3}{63}$$

Choose the answer that goes in the blank:

$\frac{3}{63}$ is ———— .

○ too small
○ correct
○ too big

✔
Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{7}{35} + \frac{2}{5} = \frac{\boxed{}}{\boxed{}}$$
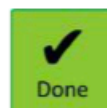
✔
Done

You may use this area for scratch work:

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{1}{6} + \frac{2}{7} = \frac{\boxed{}}{\boxed{}}$$

✔
Done

You may use this area for scratch work:

This addition is true. The questions below ask about the numbers in the equation.

$$\frac{11}{36} + \frac{9}{25} = \frac{599}{900}$$

Is $\frac{599}{900}$ greater than $\frac{11}{36}$?	Is $\frac{599}{900}$ greater than $\frac{9}{25}$?
○ Yes	○ Yes
○ No	○ No
○ Not enough information	○ Not enough information

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{1}{10} + \frac{1}{8} = \frac{\boxed{}}{\boxed{}}$$

✔ Done

You may use this area for scratch work:

Do NOT add these fractions. > means "is greater than".

Is this statement true or false?

$$\frac{19}{37} + \frac{36}{58} > \frac{36}{58}$$

○ True
○ False

✔ Done

153

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{4}{9} \; + \; \frac{5}{18} \; = \; \frac{}{}$$

✔ Done

You may use this area for scratch work:

Is this correct?

Choose the answer that goes in the blank:

is _____ .
- ○ too small
- ○ correct
- ○ too big

✔ Done

Is this correct?

$$\frac{2}{10} + \frac{2}{3} = \frac{26}{30}$$

Choose the answer that goes in the blank:

$\frac{26}{30}$ is ____ .

- ○ too small
- ○ correct
- ○ too big

✔ Done

Do NOT add these fractions. > means "is greater than".

Is this statement true or false?
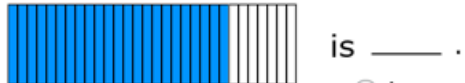
$$\frac{31}{56} + \frac{17}{42} > \frac{31}{56}$$

- ○ True
- ○ False
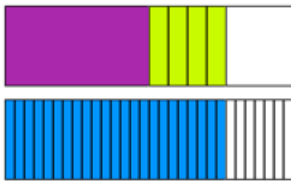
✔ Done

Is this correct?

Choose the answer that goes in the blank:

 is ____ .

- ○ too small
- ○ correct
- ○ too big

✔ Done
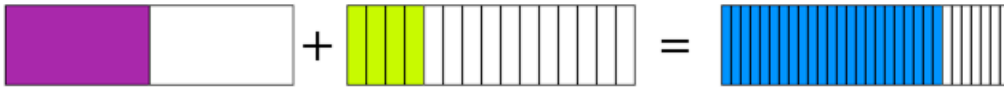
Is this correct?

$$\frac{1}{6} + \frac{3}{8} = \frac{4}{14}$$

Choose the answer that goes in the blank:

$\frac{4}{14}$ is ——— .

- ○ too small
- ○ correct
- ○ too big

✔ Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{4}{13} + \frac{1}{2} = \frac{\boxed{}}{\boxed{}}$$

✔ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{3}{8} \quad + \quad \frac{2}{6} \quad = \quad \frac{}{}$$

✓ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers
in the numerator and denominator.

$$\frac{1}{4} \quad + \quad \frac{2}{5} \quad = \quad \frac{}{}$$

✓ Done

You may use this area for scratch work:

Add. Your answer must use whole numbers in the numerator and denominator.

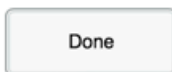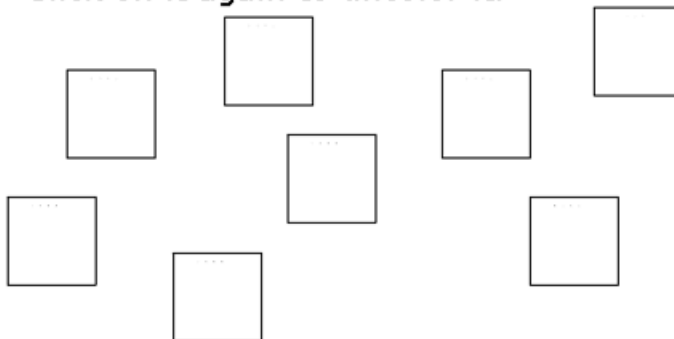$$\frac{2}{17} + \frac{6}{17} = \frac{\boxed{}}{\boxed{}}$$ ✔ Done

You may use this area for scratch work:

There are 8 blocks. Color in $\frac{1}{4}$ of them.

To color a block, click on it.
Click on it again to uncolor it.

Done

This addition is true. The questions below ask about the numbers in the equation.

$$\frac{17}{51} + \frac{13}{78} = \frac{23}{46}$$

Is $\frac{23}{46}$ greater than $\frac{17}{51}$?	Is $\frac{23}{46}$ greater than $\frac{13}{78}$?
○ Yes	○ Yes
○ No	○ No
○ Not enough information	○ Not enough information

✔
Done

A package of cookies has 9 cookies in it.

Sasha has $\frac{2}{9}$ of a package.   Malia has $\frac{6}{9}$ of a package.

What fraction of a package of cookies do they have all together?

$$\frac{\boxed{}}{\boxed{}}$$

Done

Add. Your answer must use whole numbers in the numerator and denominator.

$$\frac{2}{3} + \frac{3}{11} = \frac{\quad}{\quad}$$

✔
Done

You may use this area for scratch work: