Going Beyond the Problem Given: How Human Tutors Use Post-Solution Discussions to Support Transfer

Sandra Katz, David Allbritton, John Connelly

International Journal of Artificial Intelligence in Education 13 (2003) 79-116
IOS Press

Going Beyond the Problem Given: How Human Tutors Use Post-Solution Discussions to Support Transfer

Sandra Katz. *Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260 USA; e-mail: katz@pitt.edu and connelly@pitt.edu*

David Allbritton & John Connelly. *Department of Psychology, DePaul University, 2219 N. Kenmore Ave.,Chicago, IL 60614 USA; e-mail: dallbrit@depaul.edu*

Abstract. Two studies investigated the role and effectiveness of post-solution, reflective dialogues in physics tutorials. The first study investigated the instructional roles of post-solution discussions, their relationship to problem-solving discussions, and features that predict learning. Seven tutors individually guided 15 students as they worked on problems in the Andes physics tutoring system. Tutors adapted the post-solution discussions to students' ability levels and their performance on the current problem. Qualitative analysis of the transcripts revealed several roles of the post-solution dialogues—most prominently, explaining conceptual knowledge and integrating this knowledge with strategic, problem-solving knowledge. The number of post-solution discussions students had with their tutor, the number of discussions that abstracted from the current problem, and the number of tutor-initiated discussions predicted transfer, as measured by pre-test to post-test gain score on problems similar to those solved in Andes. Several tutorial strategies that are distributed between problem solving and post-solution reflection were identified. A framework for describing distributed plans for reflection is proposed based on these analyses.

The second study investigated whether reflection questions such as those asked by the tutors in the first study lead to better conceptual understanding and problem-solving ability, as measured by overall gain scores and gain scores on conceptual and quantitative questions. It also examined whether human tutor-provided feedback on students' responses—with its often multi-exchange, dialectic character—is more effective than a single, canned explanation. Forty-six students solved problems in Andes in one of three conditions: with no reflection questions after problem solving, with reflection questions discussed with human tutors, or with the same reflection questions followed by canned feedback (without a human tutor). Students learned more with reflection questions and feedback than without, but the canned feedback and human tutored conditions did not differ significantly. Hence, overall, these studies support the practice of implementing post-solution reflective activities in intelligent tutoring systems, but call into question the need for natural-language processing techniques to support these activities.

INTRODUCTION

Several studies have shown that one-on-one human tutoring is a highly adaptive, dynamic process. Instruction takes place mainly in response to the errors that students make while solving a problem (e.g., McArthur, Stasz, & Zmuidzinas, 1990; Person, Graesser, Kreuz, Pomeroy, & the Tutoring Research Group, 2001; Putnam, 1987). We have been studying human tutoring in two domains, avionics and physics, and have observed that tutorial discussions often continue after a problem has been solved (Katz & Allbritton, 2002; Katz, O'Donnell, & Kay, 2000). During these discussions, the tutor typically guides the student in a reflective discussion about the problem and the student's solution, clarifying concepts needed to solve the problem, explaining how the student could have solved the problem more efficiently, offering tactical advice (e.g., "break the problem down into small steps"), and reifying problem schemata

("ah yes…another pulley problem!"). As such, post-solution discussions give tutors a "second chance" to adapt instruction to the needs of individual learners.

Over the years, several developers of intelligent tutoring systems (ITSs) have incorporated post-solution, reflective tools and activities into their system. The most common tool is a reification of the student's solution trace, with or without feedback from an automated "coach" (e.g., Akhras & Self, 2000; Brown, 1985; Katz et al., 1998; Pioch, Roberts, & Zeltzer, 1997). Visual reifications of students' solution process as in ALGEBRALAND (Brown, 1985), and visual and verbal "walk-throughs" such as the Reflective Follow-up module in Sherlock 2 (Lesgold, Katz, Greenberg, Hughes, & Eggan, 1992; Katz et al., 1998) realize the traditional view of reflection as a "looking back" at the process of accomplishing a task, with critical evaluation and elaboration on that process (Dewey, 1933). Indeed, Piaget (1976) considered the ability to engage in critical reflection as one of the most advanced stages of cognitive development. White and Frederiksen (1998) support critical reflection on students' inquiry process in their ThinkerTools curriculum by offering sets of assessment criteria (called the Reflective Assessment Process) as a tool that students can use to evaluate their investigations, individually or with peers.

Research on the effectiveness of reifications of students' solution processes (Foss, 1987a, 1987b; cited in Wenger, 1987) and assessment criteria (White & Frederiksen, 1998; White, Shimoda, & Frederiksen, 1999) have shown that these reflective tools improve self-assessment and correction skills and enhance performance on subsequent tasks. Derry and Lesgold (1996) suggest that there could be a motivational as well as a cognitive explanation for the effectiveness of post-solution reflective activities such as these. Students are motivated to attend to the tutor's feedback because they just experienced impasses and realize that there is a need for some learning. Post-solution conversations can tune procedural knowledge and elaborate on conceptual knowledge, while anchoring these lessons in students' experience (Bransford, Sherwood, Hasselbring, Kinzer, & Williams, 1990).

Although there is evidence that some tools and activities to support reflection after learning tasks are effective, there is little research on reflective discussions between a human tutor and (a) student(s) that can justify or guide the implementation of other reflective activities, including natural-language post-solution conversations. This paper describes two studies that take a step towards filling this gap by addressing the following questions about post-solution, reflective dialogues in live physics tutorials:

> What instructional roles do these dialogues perform and what is the relative frequency of these roles? For example, is conceptual knowledge in focus more than strategic, problem-solving knowledge (or the reverse)?
>
> Is there evidence that human tutors adapt post-solution discussions to the needs of individual learners? If so, how is adaptive reflection carried out?
>
> How are these dialogues structured? Do they follow the chronological structure of solution traces implemented in several ITSs? Are extended exchanges (e.g., the "directed lines of reasoning" of Hume, Michael, Rovick, & Evens, 1996) common, or are simpler exchanges (e.g., Question→Response→Acknowledgement) the norm?
>
> What relationships (if any) exist between tutorial discussions that take place during and after problem solving? Are there tutorial plans that span these two phases of instruction? If so, how can distributed plans for reflection be specified so as to inform the design of reflective planners in intelligent tutoring systems?
>
> Do post-solution discussions between human tutors and students enhance conceptual understanding and transfer—that is, the student's ability to apply concepts and adapt familiar solution strategies to unfamiliar problems? If so, what features of these dialogues predict learning?

Prior research on post-solution reflection in live tutorials has contributed to our understanding of its instructional roles, discourse structure, and effectiveness. Rosé (1997) identified common types of student-tutor interactions during reflective dialogues in avionics, such as discussions of the reasoning behind taking certain problem-solving actions. Research by Katz et al. (2000) corroborated and extended these observations about the instructional roles of post-solution reflection in avionics training. Moore (1996) analyzed the structure of human experts' reflective explanations in avionics. Moore, Lemaire, and Rosenblum (1996) specified the ways in which students and tutors refer to prior explanations that take place during post-solution, reflective discussions. This research suggests that explanations evolve and become elaborated during post-solution conversations. Katz et al. (2000) show that a similar claim can be made about explanations that are distributed between problem solving and post-solution reflection. They found that these distributed explanations were more effective in resolving students' misconceptions about avionics than explanations that took place during problem solving alone. Smith-Jentsch, Zeisig, Acton, and McPherson (1998) also provide evidence that live post-solution discussions ("debriefs") can enhance performance, when coupled with pre-briefs on team training exercises.[1]

The picture of post-solution reflection that emerges from the research on live tutorials cited in the previous paragraph differs in several important ways from the nature of reflective activities commonly implemented in ITSs. First, chronological or narrative "traces" of students' solutions are rare in the post-solution conversations that we have observed, in avionics and physics tutorials (Katz et al., 2000; Katz & Allbritton, 2002). Instead, tutors tend to structure post-solution discussions around particular errors or "trouble spots" in the student's solution. This is one way in which adaptive reflection takes place. Second, coinciding with this focus on critical events and errors in the student's solution is an emphasis on *integrating* strategic knowledge with conceptual knowledge. The visual reifications of solutions that are implemented in some systems focus on the strategic process, although some traces are annotated with explanations about the reasoning behind taking (or not taking) certain actions and the conceptual basis for this reasoning (e.g., Katz et al., 1998; Lesgold et al., 1992; Pioch et al., 1997). Third, post-solution reflective conversations in live tutorials can be quite extensive. They consist of several student-tutor exchanges and are often dialectic in nature (Katz et al., 2000; Moore, 1996; Rosé, 1997). This is in sharp contrast to the single Tutor Feedback→Student Acknowledgement or Student Question→Tutor Explanation exchanges supported by current reflective modules. Finally, in live tutorials, post-solution explanations often elaborate on problem-solving explanations. There is a synergy between these two phases of instruction in live tutoring that is missing from ITSs.

These observations suggest that there is much work to be done before we can automate post-solution conversations like those that take place during live tutorials. The two studies discussed in this paper were undertaken to further our understanding of the instructional roles of live post-solution discussions, the nature of adaptive reflection and distributed reflective discussions, and how these discussions support learning. In the first study, we investigated the roles of post-solution discussions in physics tutorials. Seven tutors individually guided 15 students who worked on problems in the Andes physics tutoring system (e.g., Gertner & VanLehn, 2000; see also http://www.pitt.edu/~vanlehn/andes.htm). We analyzed the relative frequencies of various post-solution roles and relationships between problem-solving and post-solution discussions. We then considered whether the post-solution discussions were tailored to students' overall ability and performance on the current problem—for example, whether tutors tended to favor certain post-solution roles and informational relationships for higher-ability students than lower-ability students. Next we investigated features of the post-practice discussions that predict learning, as measured by pre-test to post-test gain score. Finally, we took a closer look at several reflective discussions that were distributed between problem solving and post-practice reflection, in order to

[1] This research did not isolate the effects of the de-briefs from pre-briefs or other elements of the team training program

determine if there were recurring distributed strategies for reflection. Having identified several distributed strategies, we developed a framework to describe them in a way that can guide automated planning of reflective dialogues.

The main goal of the second study was to do a controlled evaluation of some of the questions that tutors posed to students in the first study, and hence to assess the value of incorporating these types of questions in automated reflective modules. Borrowing terminology from Lee and Hutchison (1998), we refer to these questions as "reflection questions." Lee and Hutchison (1998) found that reflection questions after worked examples enhanced students' problem-solving ability in balancing chemistry equations. The second study investigated whether reflection questions and feedback on students' responses would correlate with pre- to post-test gain scores in conceptual knowledge, quantitative problem-solving ability, both, or neither. As such, this study extends Lee and Hutchison's work by examining the effectiveness of reflection questions after problem solving (as opposed to example studying), with respect to both conceptual understanding and problem-solving ability.

Because both studies were situated in Andes, we first provide an overview of this system. We then discuss the studies and their results in turn and conclude with a summary of how post-solution reflection can enhance the ability of human and automated tutors to "care for learners."

OVERVIEW OF THE ANDES TUTORING SYSTEM

We describe Andes from the user's perspective, as the tutors and students in our studies saw it. Other papers provide more detail on the Andes design principles and architecture, coaching modules, student modeling engine, and evaluation.[2]

Andes is a collaborative effort between the University of Pittsburgh and the U.S. Naval Academy. It was developed primarily to be used as a "homework helper," by students in the U.S. Naval Academy and the U.S. Department of Defense Dependent Schools. The system allows students to work on a large number of practice problems in classical Newtonian physics, with help from Andes.

An example of the Andes interface, as the student would see it when selecting one of these exercises, is shown in Figure 1. Students read the problem statement, sketch the physical situation (e.g., draw a "free body diagram"), and record their answer in the top-left pane; they define variables in the top right pane; they enter solution steps in the bottom right pane, which is intended to look like a piece of paper; and they receive hints in the lower left pane. Although most coaching that students receive in a typical session is on demand, Andes occasionally initiates "mini-lessons" when its student modeling engine identifies particular concepts that the student appears to lack or misunderstand.[3] Andes also gives immediate feedback on the correctness of all student actions—forces and vectors included in diagrams, variable definitions, and equations entered on the scratchpad—using a red/green highlighting convention. This is intended to maximize learning opportunities and minimize the time spent going down incorrect solution paths.

If students cannot figure out why a step or drawing object was marked wrong (red) or reach an impasse, they can ask for help. Andes provides conceptual and procedural help that is designed to encourage students to think on their own. It does this by giving general feedback first and becoming increasingly directive if the student remains stuck. For example, a first-level hint on a faulty equation

[2] See the reference s cited in http://www.pitt.edu/~vanlehn/andes.htm.

[3] The mini-lesson utility was not yet implemented in the version of Andes used in our experiments. Hence, most coaching messages were solicited by the student. Several "warning-messages" were unsolicited. These messages mainly pertained to interface usage; for example, Andes advised the student to choose a different variable name, if the one the student selected was already in use. A handful of warning messages addressed physical concepts (e.g., "Average Speed occurs during a time interval, not at a point in time")

might point out a problem with one of its variables ("Think about the direction of 'force-on-driver'"); the most directive, "bottom-out" hint would display the correct equation.

Students proceed in this fashion—drawing diagrams, defining variables, writing equations, and soliciting help as needed—until they solve the problem. A correct solution is verified by green feedback on the student's final equation, which specifies the sought-after quantity.

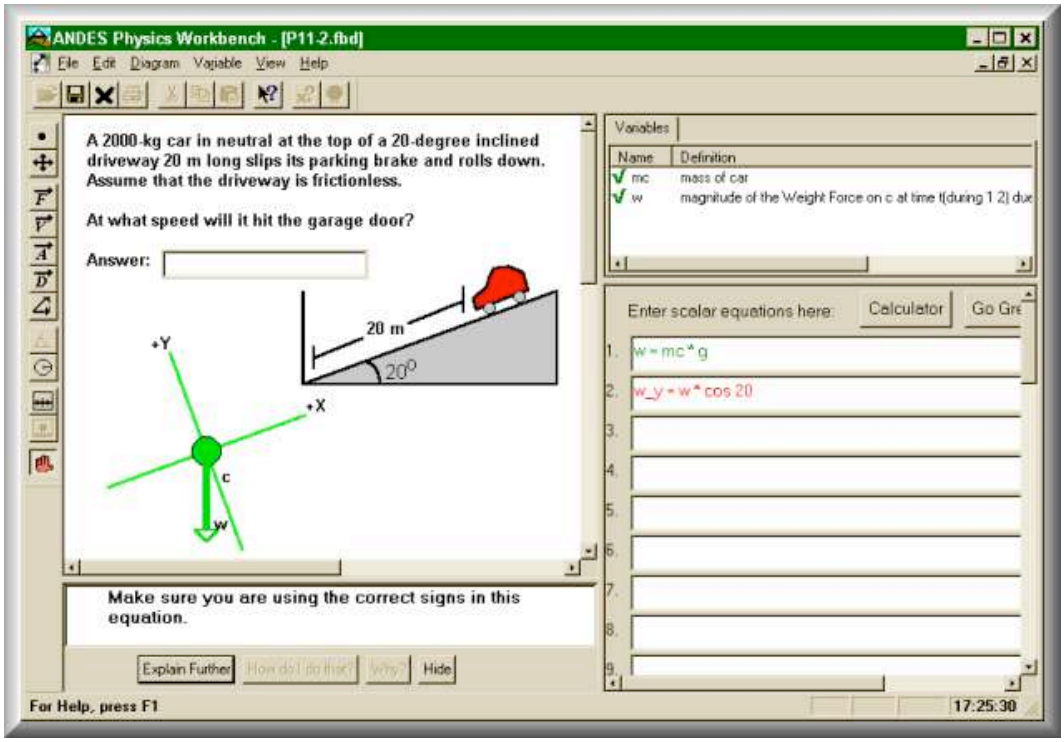


Figure 1. Example of the Andes interface

A STUDY OF POST-SOLUTION DISCUSSIONS IN PHYSICS TUTORIALS

In prior studies of avionics tutoring, Katz et al. (2000) observed that problem-solving dialogues tended to focus on solving the task at hand. Tutors advised students on the actions to take next and justified their advice in terms of a global plan and the local goals needed to carry out this plan. However, there was very little discussion during problem solving about how the electronic system worked or about general "tricks of the trade" for troubleshooting electronic faults. These conceptual and tactical discussions tended to be deferred until the post-solution debrief.

The study described in this section was conducted in part as a first test of the generality of these findings, by investigating human tutoring in a different domain. We chose physics—in particular, elementary mechanics—because, as in avionics, problem-solving ability depends upon conceptual understanding, at least for more difficult problems.[4] We were curious to see whether strategic and

[4] It is sometimes possible to solve mechanics problems using scripted methods and little understanding, especially in the early stages of instruction.

procedural discussions would tend to dominate problem solving while discussions that abstracted solution schemata, tactics, and domain principles would be the province of post-solution conversations, as we observed in the avionics tutoring corpus. In addition to investigating the roles of post-solution discussions in a new domain and the relative frequency of these roles, this study considered whether these discussions are adaptive, how they relate to problem-solving discussions, whether they enhance conceptual understanding and transfer, and (if so) which dialogue features predict learning.

Method

Data Collection

Fifteen students who were currently enrolled in an introductory physics course at the University of Pittsburgh volunteered to participate in the study. They were paid a nominal amount. Each student was randomly assigned to one of seven human tutors, also paid volunteers. Tutors had prior experience teaching physics in a classroom or one-on-one tutoring setting; some had done both. Since there were fewer tutors than students, some tutors worked with more than one student. Tutor 1 worked with five students, Tutor 2 with three students, Tutors 3 and 4 with two students; the remaining tutors (Tutors 5, 6, and 7) each worked with one student. No students knew their tutor prior to the study.

Students first completed a background questionnaire. This provided us with information about which introductory physics course they were taking (algebra-based or calculus-based), their Scholastic Assessment Test (SAT) Math and Verbal scores, and whether or not they had taken physics in high school. Students then took a 50-item pre-test. Of the 50 questions, 37 were designed to measure students' ability to solve quantitative mechanics problems, while the remaining 13 were designed to measure students' conceptual understanding of mechanics. At this point, students were given instruction in how to use Andes, the intelligent tutoring system that would serve as both a problem-solving environment and communication medium for interacting with their tutor.

After the Andes orientation, students worked on 24 problems in the tutoring system, over the course of three to five sessions. Each session lasted 2 to 4 hours. The tutor and student sat in separate rooms and interacted via teletype. Andes automatically logged students' actions and their conversations with the tutor. Andes' automated coaching was suppressed during the experiment so that all of the help that students received came from the live tutors.

In order to highlight the roles of post-solution dialogues and the potential effect of prohibiting these discussions on problem-solving interactions, we assigned problems to one of two possible formats, "debrief" or "no-debrief." At the start of each problem, the experimenter told the student and tutor whether or not they would be allowed to discuss the problem further after the student solved it. There were 12 problems in each of these within-subject conditions—that is, 12 were debrief and 12 were no-debrief. Each student solved the same set of 24 problems in the same order, but there were two patterns with respect to session format. For example, in one pattern, the first problem was in debrief format and in the other pattern the first problem was in no-debrief format. Students were randomly assigned to work through the problems in one of these two patterns. The patterns were designed in such a way that participants could not guess the format of the next problem. Two patterns were used so that tutors who worked with more than one student could not learn the order of debrief and no-debrief problems.

Students took a post-test after they completed the problems. The post-test was the same as the pre-test. Students then filled out a questionnaire that asked: (a) how the student and his or her tutor used the post-solution discussions—that is, what sorts of topics, if any, they discussed; (b) whether they found these discussions beneficial; and (c) if and how they interacted differently with their tutor when they were not allowed to discuss the problem after having solved it. Tutors who worked with more than one student

completed the questionnaire once, after all of their students had finished. They were asked to consider each of the students they worked with when answering the questions. Had we asked tutors to fill out a separate questionnaire immediately after each of their students finished, the tutors' behavior with subsequent students might have been biased because they would have known what sorts of questions we were addressing in the study.[5]

Dialogue Analysis

The data corpus consists of 310 transcripts, 160 from debrief problems and 150 from no-debrief problems.[6] The analyses reported in this paper focus on the transcripts for problems that allowed students and tutors to discuss a problem after the student solved it (i.e., the debrief problems).

Participants did not always take advantage of the opportunity to discuss a problem further. Among the 160 debrief problems, 101 (63%) contained at least some post-solution discussion. The amount of discussion varied from a single comment (e.g., "that was a tough problem") to nearly 100 dialogue turns. Among the remaining 59 debrief sessions, 34 (21%) contained evidence that the participants were aware that they could discuss the problem further but chose not to. In these sessions, the tutor typically prompted the student to ask questions (e.g., "Any questions?") and the student (predictably) replied "No." The other 25 debrief sessions (16%) left us wondering whether the student and tutor knew what the session format was, because they simply ended their discussion when the student entered his or her answer into the Andes "Answer" box. In a few cases, there was direct evidence of confusion about session format (e.g., when a student asked his tutor, "can we still talk about this?" after solving the problem).

We marked the start of the post-solution dialogues at the point where the student got the correct answer and the human tutor confirmed it. Typically this was followed by the tutor soliciting questions from the student, or by the student initiating a question. We segmented the post-solution dialogues into sub-dialogues, one sub-dialogue per topic and its associated sub-topics. Table 1 presents a sample post-solution discussion with two sub-dialogues:[7] one that the tutor initiates to address the student's tendency to do math before she writes an equation symbolically (turn 4), and a second that the student initiates about the difference between uniform circular motion and rotation, which she mistakenly refers to as "rolling" (turn 5). The first sub-dialogue extends a problem-solving discussion in which the tutor cautioned the student not to get ahead of herself. Note that a sub-dialogue may itself contain several sub-topics. For example, the second sub-dialogue in Table 1 contains an embedded discussion about coordinate systems, which the student initiates in a follow-up question (turns 13-19). However, we coded two levels of discourse—dialogue and sub-dialogue—because the features we were interested in (specified below) were identifiable at these levels. Among the 101 sessions with post-solution dialogues, 78 (77%) contained a single dialogue, while the remaining 23 (23%) contained one or more sub-dialogues. For simplicity, we will refer to both single dialogues and sub-dialogues as "post-solution dialogues." The corpus consists of 146 coded post-solution dialogues.

To address the questions stated in the introduction to this section, we coded the following features of post-solution dialogues:

---

[5] In retrospect, it would have been better to ask the four tutors who worked with more than one student to complete a separate questionnaire for each student, after all of their students had finished, instead of asking them to fill out one questionnaire representing all students. Although these tutors commented about each student in some questions, they did not do this systematically.

[6] Due to time constraints and other factors, four students partially completed the 24-problem set before taking the post-test.

[7] Typos and spelling errors have been edited for clarity, in Tables 1, 2, 7, 8, and 10.

1. **Initiator:** Who initiated the dialogue: the tutor, or the student? We tagged a turn as an initiation only if it led to an instructional dialogue. For example, if the tutor asked the generic "any questions?" and the student then asked a question, we coded the student as the initiator, not the tutor. This point is illustrated in Table 1. Although the tutor asked the student if she had "any questions" in sub-dialogue 1 (turn 2), the student did not take up this request until after the tutor commented on the student "second-guessing" herself (turn 5). The initiator of the first sub-dialogue is the tutor; the initiator of the second sub-dialogue is the student.

Table 1. A Post-Solution Discussion With Two Sub-Dialogues

Sub-dialogue 1

1. S: I got 13.41 m/s
2. T: yes that is what I got...any questions on this ?
3. S: I don't think so
4. T: ok...you seem to be second guessing yourself...don't do that! especially on an exam...if you know the theory well then the math is really trivial (so far) and I know you know how to do that

Sub-dialogue 2

5. S: I have a question----this is uniform circular motion, briefly what is the main difference between this and rolling is it because position is being changed with respect to some other object?
6. T: well…if you mean rotation? like pure rotation ?
7. S: yes
8. T: then you are talking about a rigid body being rotated about some fixed axis...
9. S: ok
10. T: in that case the entire variables are different
11. S: right
12. T: our acceleration is now alpha which is a change in omega over time and w is now defined as a change in theta over time
13. S: is it like changing coordinate systems?
14. T: yes think of a circle
15. S: okay
16. T: how do you relate linear coordinates to circular coordinates... s=r(theta) right ?
17. S: yes
18. T: in this we still have s...we are only changing direction of velocity...the magnitude is constant
19. S: okay that makes sense, I see

2. **Instructional role:** What type of knowledge was the tutor trying to teach to the student during the post-solution dialogues: strategic knowledge (e.g., an abstract schema that could be applied to similar problems), physical concepts or principles, general problem-solving tactics, and so forth? What other goals (besides knowledge enhancement) was the tutor trying to achieve—increasing the student's confidence or motivation, assessing the student's performance on a particular

problem or on a class of problems, session management (e.g., interface usage) matters, and so forth?

3. **Information status:** Does the post-solution dialogue expand upon a discussion that took place during problem solving, summarize it, or bring new issues to the table?

Below we specify our coding schemes for instructional role and information status and discuss how they were developed. Marking these features allowed us to analyze the relative frequency of various instructional roles in this corpus of post-solution dialogues, determine whether there was evidence that tutors tailored these discussions to students' individual needs, identify features of post-solution discussions that correlate with pre-test to post-test gain scores, and identify tutorial plans that span the two phases of instruction.

Instructional role. The comments that students and tutors recorded on the post-participation questionnaire served as a good starting point for identifying the instructional roles of post-solution discussion in this domain. As shown in Table 2, most participants described the conceptual focus of their post-solution dialogues and the role these dialogues played in developing students' strategic, solution-planning skills. To a lesser extent, students commented on the dialogues' tactical role—that is, learning general problem-solving tips such as "break a problem down into small steps," "don't worry about the math until you express the equation symbolically," and so forth. As Table 2 demonstrates, students' comments on the roles of the post-solution dialogues tended to be consistent with those of their tutors.

Table 2. Students' and Tutors' Retrospective Comments on the Post-Solution Dialogue

Representative Student Comments	**Corresponding Comments from their Tutor**
1. Sometimes my tutor would add a "twist" to the problem—especially if I solved very quickly. Also, I could ask about part of the problem or theory I was unsure of. She asked questions to ensure I knew the "why" of something.	1. I usually tried to give another example of the problem. Often, the students would ask questions on concepts.
2. We discussed other ways of solving the problem.	2. A few times the student had some general questions about the material which we discussed during the debrief time. There were times when the student did the problem rather quickly and got the right answer but really did not have the best approach. So a couple of times I commented about that and hinted at some other way of doing the problem.

Representative Student Comments	**Corresponding Comments from their Tutor**
3. My tutor first asked if I had any questions and that would lead us to discussion of the principles used and extending these to other solutions. He would ask, for example, how a problem would change if there was no gravity.	3. We generally talked specifically about the problem that was just completed. But I do recall conversations dealing with general problem solving techniques and discussing how I may have gone about solving a particular problem. I also used this time to talk about certain problems that seemed to be dealing with a certain topic but the "real problem" may be something else and the student needs to be careful by reading the problem more than one time.
4. The time was mainly spent on problem solving strategies, sometimes referring to the previous problem, sometimes more general concepts. Often, if I was having trouble with the problem in question my tutor would basically restate the concept in question and made sure I understood it completely before continuing.	4. I usually like to bring up issues and questions as the student is working. I will, however, try to sum up the general physical issues at the end, or reinforce a correction to an earlier misconception.
5. I just asked questions to understand the wording of the question and be able to identify the information needed to solve the problem and ignore extraneous information.	5. My general comments about “debrief” discussions are that they are more for the tutor’s benefit than the student. The student has just worked a problem he/she has never seen before, and depending how much “physics smarts” he/she has, what might he/she want to say?
6. We used “debriefing” sometimes to relate the problem to other situations where it would hold true or he would sometimes ask me questions to see if I really understood what was covered in the problem.	6. I usually used this time to rehash points that were made before. … I feel that the student understands the points that I’m trying to make better if it's at the time that the problem is occurring, especially since the only way that we could communicate was through typing to each other. There were a couple of times that I tried to put this off until the end, but the conversation didn't go as smoothly.
7. We typically didn’t use them for much other than comments about the difficulty (e.g., “Wow! That was hard”) or chit chat (e.g., “So, how are you doing?”).	

A closer examination of the data allowed us to further specify the instructional roles that participants mentioned and to identify other roles. The resulting scheme describes dialogue roles in terms of the type of domain knowledge focused upon: *conceptual*, *strategic*, *procedural*, and *tactical*. We also identified a fifth, *general* category that represents other goals besides instruction in domain knowledge (e.g., motivate the student, bolster his or her confidence, assess the student's ability to solve mechanics problems). Each post-solution dialogue could be tagged with more than one role descriptor. For example, some dialogues contained discussions of conceptual knowledge along with motivational comments; some dialogues integrated conceptual and strategic knowledge. The relative frequencies of these roles will be discussed in the Results and Discussion section.

Conceptual Knowledge Dialogues: Focus on physics concepts and principles (e.g., energy, Newton's laws of motion). There are six sub-categories of conceptual knowledge dialogues. Because we wanted to understand what participants meant when they said that the post-solution discussions focused on concepts, we coded the dialogues in terms of these sub-category labels:

Conceptual generalization: Help the student understand concepts or principles illustrated in the current problem and how these concepts apply to various physical situations. A common tactic for doing this is to present a "what if" scenario, a variation on the current problem. For example, if a problem applies Newton's Second Law (NSL) to a stationary object, the tutor might vary the problem so that the object is moving and discuss how the same law applies to this situation. Conceptual generalization is exemplified in student and tutor comments 1, 3, and 4 in Table 2.

Conceptual specialization: Clarify the distinction between related concepts—for example, the difference between instantaneous, average, and constant acceleration.

Correct knowledge gap: Explain a concept or state a piece of declarative knowledge that the student apparently lacks—for example, what the unit "Newton" means; what NSL says about the relation between the forces on an object and its acceleration. This is often done to resolve a misconception.

Correct misconception: Correct a piece of faulty knowledge or a flawed mental model (e.g., that all force problems are static and, therefore, the net force is always equal to zero). Correct misconception is illustrated by tutor comment 5 in Table 2.

Check understanding of correct solution: If a student solved the problem perfectly, except perhaps for arithmetic errors, determine whether the student understood the physics behind the problem and did not merely solve it by rote. This is often done by posing "what if" scenarios and Socratic-style dialogues. Tutor comment 3 in Table 2 expresses the latter.

Verify previous instruction: If a topic was discussed during problem solving, check that the student understood the points covered in that discussion.

Strategic Knowledge Dialogues: Focus on planning—global planning of goals, and local planning of sub-goals and the actions needed to achieve them. There are two sub-categories of strategic dialogues:

Strategic generalization: As with conceptual generalization, go beyond the case at hand; help the student realize that the strategy used in the current problem applies to a whole class of

problems—for example, all force problems involving a static object, all "work done by a variable force" problems, all energy conservation problems. Similarly, clarify how a particular schema applies to the current problem. Strategic generalization is captured by student comment 6 in Table 2.

Alternative strategies: Teach different strategies for solving the same problem or for achieving a particular solution step, or help the student understand why one strategy is preferable or equivalent to another. This role is represented in student/tutor comments 2 and 4 of Table 2.

Procedural Knowledge Dialogues: Focus on the correct execution of actions taken to achieve goals. In this domain, this typically means calculation—for example, which trigonometric function to use, unit errors, sign errors,[8] violations of domain standards (e.g., number of significant figures). There are two sub-categories of procedural knowledge dialogues:

Standards: Discuss the correct use of units and domain conventions such as rounding off to the number of significant figures. For example, during a post-solution dialogue, a tutor advised his student that her answer "looks like the national debt."

Discuss alternative procedure: Advise the student about a more efficient or "cleaner" way to carry out a goal—for example, a different kinematics equation to use, instead of one that requires solving a quadratic equation.

Tactical Knowledge Dialogues: Teach "tricks of the trade" for solving quantitative problems, such as breaking problems into more manageable sub-goals, reading the problem statement carefully, and expressing relations symbolically before instantiating variables. Tactical knowledge is referred to in student comment 5 and tutor comment 4 of Table 2.

General Dialogues: Focus on goals other than instruction in domain knowledge. There are five sub-categories of general dialogues:

Assess performance on current problem: This includes general evaluative comments like "Great!" or "much better this time," and more targeted remarks like, "it seems like you did a lot of extra work."

Assess general performance: Comment on the student's overall ability in physics, or in solving a class of problems (e.g., energy problems, kinematics problems).

Motivate student: Comment on the benefits of understanding physics (e.g., "you need to know how the world works") and the benefits of following the tutor's advice and listening to his or her explanations, like "I'm sure [your professor] will ask what the difference is [between conservative and non-conservative energy] on your next test."

Comment on problem: Comment on the problem's difficulty or instructional value (e.g., "these problems are designed to train you to use your intuitions and ignore useless information").

[8] Sign errors often indicate conceptual knowledge gaps or misconceptions, rather than math slips. For example, the student may not understand the vector relationships in a problem. We code these instances using conceptual knowledge sub-categories (i.e., correct knowledge gap, correct misconception).

Session management: Establish communication protocols, such as "write notes to me in the Comments Box, not in the equation lines," "define your variables so that I know what they mean."

Information status. Analyzing post-solution dialogues from the perspective of the instructional roles that we have described can be misleading. It is tempting to view these dialogues as though they exist in isolation, apart from any previous conversations. However, as the comments in Table 2 suggest, students and tutors view the two phases of instruction in relation to each other. This is evident by terms such as "sum up," "reinforce," and "give another example." In this section, we describe our system for coding these relationships.

A post-solution dialogue's status can be *new* or *old*. *New* dialogues address topics that were not discussed during problem solving, while *old* dialogues summarize or elaborate on problem-solving discussions. We sub-divide these broad categories into four information status descriptors, which are described in Table 3.

Table 3. Information Status Categories

"**Old" Dialogues:**
Elaboration: Elaborated dialogues add information to a problem-solving discussion.[9]
Restatement: Restated dialogues have a precursor during problem solving. However, they do not add information; they merely summarize or restate what came before.
"New" Dialogues:
New/domain-based: These dialogues are unrelated to any discussion that took place during problem solving. The subject matter focuses on the domain (e.g., elementary mechanics).
New/general: These dialogues are also unrelated to any discussion that took place during problem solving but their content is not domain-based. Instead, they deal with assessing performance, motivating the student, bolstering confidence, session management, and so forth.

Results and Discussion

Two people coded the dialogue features described in the preceding section. Fifteen transcripts were selected at random to test for inter-rater reliability. Thirteen transcripts contained a post-solution conversation, and the coders' judgments of whether a debrief occurred were in perfect agreement (100%). The coders agreed on both the number of sub-dialogues and the sub-dialogue boundary locations for 12 of the 13 debrief transcripts (92%). Agreement rates were 92% for initiator (kappa = .83), 85% for information status (kappa = .78), and 94% for instructional role (kappa = .77).

[9] If a tutor stated something in a more generic way during the post-practice dialogue than during problem solving, we coded the associated dialogue as an elaboration because the more generic form implies that the information is broadly applicable, not just relevant to the problem at hand. For example, if a tutor advised the student to "Start with Newton's second law" during problem solving, and said "Whenever you see forces, think of Newton's second law" during the post-practice discussion, we coded the associated sub-dialogue as an elaboration and the instructional role as "strategic generalization."

Frequency Analyses

Overall, approximately 34% of the dialogues were initiated by students, 66% by their tutors. Figure 2 shows the relative frequency of student-initiated and tutor-initiated dialogues with respect to instructional role. Figure 3 does the same with respect to information status. Tutors initiated more dialogues than students in all categories, for both instructional role and information status. However, the fact that nearly a third of the post-solution dialogues were initiated by students is worth noting, given that studies of human tutoring during problem solving have found that students ask few questions (e.g., Graesser, Person, & Magliano, 1995). Perhaps post-solution dialogues encourage active learning.

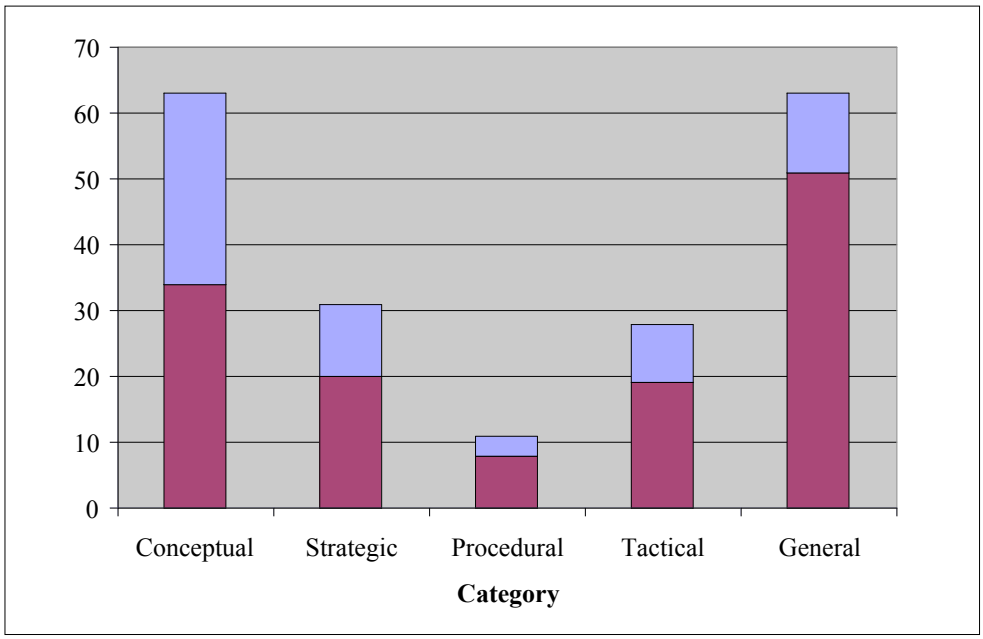


Figure 2. Frequency of instructional role categories

As Figure 2 reveals, most post-solution dialogues were conceptual or general (63 dialogues), followed by strategic (31). (Recall that dialogues could be coded with more than one role descriptor, so numbers do not add up to 146.) The high frequency of conceptual dialogues supports tutors' and students' retrospective comments about how they used the debrief sessions, as illustrated in Table 2. However, the high frequency of elaborations shown in Figure 3 hints at the possibility that conceptual discussions were not typically deferred until debrief, as they were in the avionics corpus.

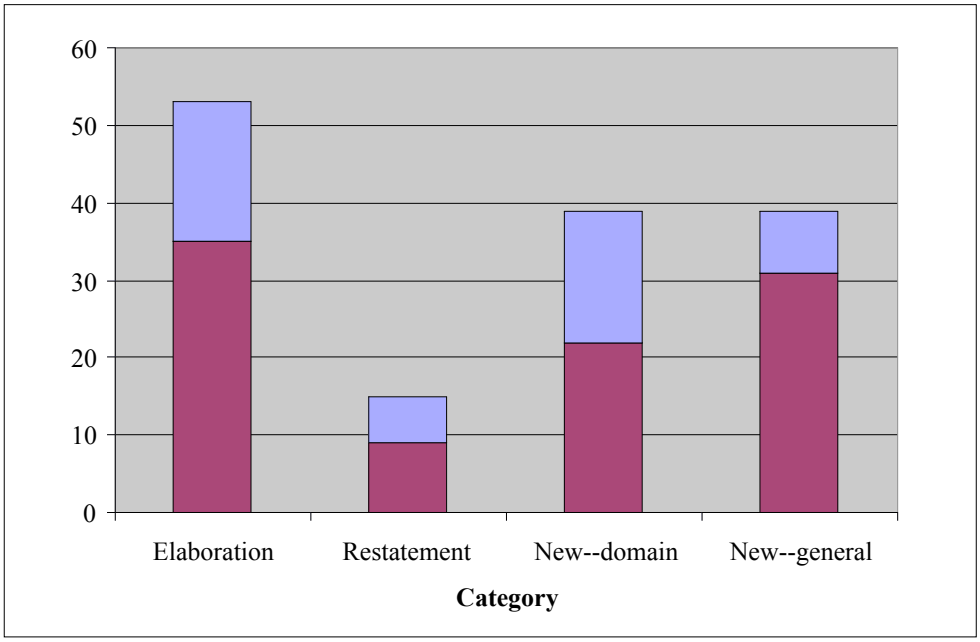


Figure 3. Frequency of information status categories

Indeed, the relationships between information status and instructional role shown in Figure 4 support this observation. In total, 68 dialogues were coded as relating to old information (elaborations and restatements) and 78 were coded as new (new—domain and new—general). Although the overall numbers of old and new topics were not significantly different, $t(14) = -0.57$, the patterns of old versus new information differed across the five instructional roles (see Figure 4). For both strategic and tactical knowledge, there were significantly more *old* (elaborated or restated) dialogues than *new*, $t(14) = 2.18$ and $t(14) = 2.82$ respectively, $p < .05$. A similar (but non-significant) trend was observed for conceptual dialogues, with 38 *old* and 27 *new* or deferred until the post-solution discussion, $t(14) = 0.99$, *ns*. These findings support our informal observation that the tutors addressed strategic (planning, or what-to-do-next) errors as they occurred during problem solving, and supported their advice with conceptual explanations. They then used the post-solution discussion to elaborate on these explanations and abstract the solution schema applied to the current problem.

For procedural and general knowledge, however, the reverse pattern holds. There were non-significantly more *new* than *old* procedural dialogues, $t(14) = -0.62$, and significantly more *new* than *old* general knowledge dialogues, $t(14) = -2.65$, $p < .05$. This is not surprising, given that—in contrast to strategic errors—procedural errors do not always prevent the student from getting the correct answer, so perhaps tutors felt comfortable delaying discussion of them. The same holds true about most of the topics discussed in general dialogues—assessment, comments about the problem, session management issues. Problems in these areas not only can wait, but also might be distracting to talk about while the student is solving the problem.

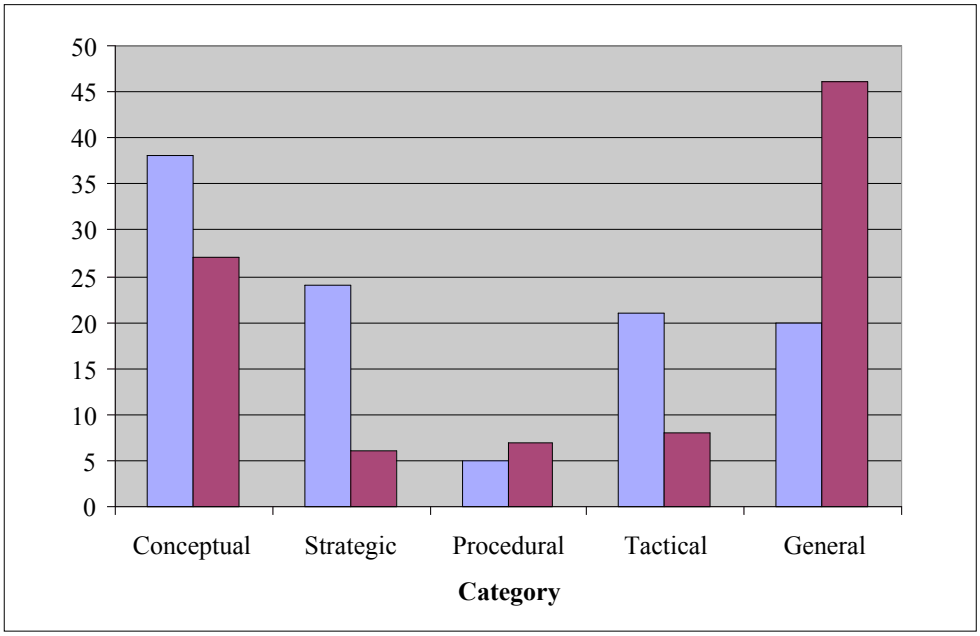


Figure 4. Relationship between information status and instructional role

Evidence of Adaptive Tutoring in the Post-solution Discussions

Tutors' strategies varied considerably from student to student, as is evident from the variability in the number of dialogues tutors initiated in each of the four information status categories (displayed in Table 4) and five instructional role categories (Table 5). The variability between students of a single tutor is at least as evident in Tables 4 and 5 as the variability between tutors.

Furthermore, tutors appeared to vary their strategies in response to differences in students' capabilities, as measured by the pre-test of mechanics that was issued before the tutoring sessions began. To find out how tutors tailored their strategies, we examined correlations between overall pre-test scores and the number of dialogues initiated by tutors that fit each information status category and instructional role category.

Table 4. Counts of Tutor-Initiated Dialogues per Information Status Category

Tutor/Student Pair	Elaboration	Restatement	New-domain	New-general
T1				
S1	4	0	5	5
S2	0	0	2	1
S3	3	1	2	3
S4	2	1	1	3
S5	9	0	0	3
T2				
S6	1	0	1	3
S7	3	2	1	0
S8	1	0	1	1
T3				
S9	3	0	1	1
S10	1	3	4	4
T4				
S11	0	1	0	3
S12	1	0	2	1
T5				
S13	2	1	0	2
T6				
S14	2	0	1	1
T7				
S15	3	0	1	0
Total Count	35	9	22	31

Table 5. Counts of Tutor-Initiated Dialogues per Instructional Role Category

Tutor/Student Pair	Conceptual	Strategic	Procedural	Tactical	General
T1					
S1	7	3	1	0	8
S2	2	1	0	0	1
S3	4	4	0	1	4
S4	3	1	0	0	4
S5	5	7	0	4	5
T2					
S6	0	0	2	0	3
S7	2	1	2	3	2
S8	1	0	0	0	2
T3					
S9	1	1	0	2	2
S10	3	1	1	4	8
T4					
S11	1	0	0	0	3
S12	0	0	1	1	2
T5					
S13	1	0	1	1	2
T6					
S14	0	0	0	3	3
T7					
S15	4	1	0	0	2
Total Count	34	20	8	19	51

Information Status. Tutors initiated more elaboration dialogues during post-solution reflection with students who had low pre-test scores than with students who had high pre-test scores (r = -.56, $p < .05$). By contrast, the number of new—domain dialogues initiated by tutors was positively correlated with pre-test scores (r = .57, $p < .05$). No other information categories correlated significantly with pre-test scores.

These correlations reflect a general tendency for tutors to initiate more dialogues reinforcing old information (restatement and elaboration categories combined) with students who had low pre-test scores, and more dialogues introducing new information (new—domain and new—general combined) with students who had higher pre-test scores. These tendencies are reflected in a negative correlation between the number of dialogues initiated by tutors about old information and student pre-test scores (r = -.36, $p <$ .05) and in a positive correlation between number of new-information dialogues and pre-test scores (r = .41, $p < .10$).

Tutor 1 provides a striking case study of this tendency to reinforce problem-solving discussions during the reflective phase with weaker students and introduce new discussions with stronger students. Student 1 had the highest pre-test score of all the students in this study (78% correct) and student 5 had the lowest pre-test score (16% correct). Tutor 1 used drastically different strategies with these two students, as can be seen in Table 4. The tutor initiated predominantly elaborative (old information) dialogues with student 5, and predominantly new—domain and new—general dialogues with student 1.

Instructional Role. Tutors initiated more strategy dialogues during the reflective phase with students who had low pre-test scores than with students who scored higher on the pre-test (r = -.58, p < .05). Although none of the other individual knowledge categories were significantly correlated with pre-test scores, there were statistically marginal trends for tutors to initiate more tactical dialogues with low-pre-test students (r = -.41, p < .10), and for tutors to initiate more procedural dialogues with high-pre-test students (r = .38, p < .10).

To better understand the pattern of the five types of knowledge categories tutors used, we entered the number of dialogues of each type that tutors initiated as predictors in a linear regression in which overall pre-test score was the criterion variable. The full regression model accounted for 91% of the variance (adjusted R^2 = .86). Tutors initiated more strategic dialogues with students who had lower pre-test scores (β = -1.18, t = -6.18, p < .001). This may be because these students needed more help with planning solutions for different types of problems. In addition, tutors initiated more conceptual (β = .76, t = 3.45, p < .01) and procedural (β = .20, t = 1.85, p < .10) dialogues with students who had higher pre-test scores. Perhaps this reflects a tendency on the tutors' part to ensure that stronger students understood why they did what they did (with respect to conceptual discussions), and how they could have achieved particular sub-goals differently (with respect to procedural discussions). The regression coefficients for number of tactical and general dialogues were not significant.

Evidence that Post-Solution Dialogues Support Learning

We measured learning in terms of *overall gain score* from the pre-test to the post-test, *qualitative gain score*—with respect to the 13 qualitative questions—and *quantitative gain score*—with respect to the 37 quantitative questions (see Footnote 14 for elaboration on our use of the gain score measure). Two raters scored the tests. Agreement was 92.6% (kappa = .84).

We predicted that gain scores in all three categories would correlate positively with the number of post-solution discussions that students had with their tutor and certain characteristics of these dialogues. In particular, we expected that post-solution dialogues that abstracted from the current problem would promote near transfer—as measured by quantitative gain scores—and conceptual understanding—as measured by qualitative gain scores. Several instructional roles capture various types of abstraction, namely those shown in Table 6. A brief definition of these categories is restated in Table 6 for convenience.

To test these predictions and investigate other characteristics of the post-solution dialogues that potentially support learning, we searched for correlations between the three measures of learning (qualitative, quantitative, and overall gain score) and the following features of the post-solution dialogues in which each student participated: (a) the total number of post-solution dialogues, (b) the number of post-solution dialogues that carried out abstraction functions (Table

Table 6. Abstraction Functions of Post-Solution Dialogues

Conceptual generalization: help the student understand concepts associated with the current problem and how these concepts apply to various physical situations.
Conceptual specialization: clarify the distinction between related concepts—for example, the difference between instantaneous, average, and constant acceleration.
Correct knowledge gap: explain a concept or state a piece of declarative knowledge that the student apparently lacks; often done to resolve a misconception.
Correct misconception: correct a piece of faulty knowledge or a flawed mental model
Strategic generalization: help the student: (a) understand that the strategy used in the current problem applies to a whole class of problems (e.g., all conservation of

energy problems), (b) recognize that a particular schema applies to the task at hand, or (c) adapt a schema to physical situations that differ in particular ways.

Alternative strategies: teach different strategies for solving the same problem or for achieving a particular solution step, or help the student understand why one strategy is preferable or equivalent to another.

Problem-solving tactics: teach "tricks of the trade" for solving quantitative problems, such as breaking complex problems into more manageable sub-goals, reading the problem statement carefully, and expressing relations symbolically before instantiating variables.

6), (c) the number of abstraction dialogues that were student-initiated and tutor-initiated, and (d) the number of dialogues tagged with each of the information status categories described in Table 3, plus the aggregated categories *new* and *old*. Four factors correlated significantly ($p < .05$, two-tailed) with overall gain score and two with quantitative gain score (we list the correlation statistics for overall gain score followed by the same for quantitative gain score): the number of dialogues that elaborated upon problem-solving dialogues ($r = .59$; $r = .59$); the number of old (elaborated and restated) dialogues ($r = .52$; $r = .48$, $p = .07$); the number of abstraction dialogues ($r = .55$; $r = .52$); and the number of abstraction dialogues that were initiated by the tutor ($r = .56$; $r = .49$, $p = .07$).[10] None of the examined factors correlated significantly with qualitative gain score.

These results suggest that post-solution dialogues support learning—in particular, near transfer of problem-solving skills similar to those used in solving the Andes problems—especially when they abstract the concepts and strategies associated with the current problem, elaborate or restate problem-solving discussions, and are initiated by the tutor.

Opportunists, Stashers, and Parcelers: Tutoring Styles With Respect to Staging Instructional Dialogues

One-on-one tutoring is a complex decision-making task, as are other forms of instruction such as mentoring and classroom teaching (e.g., Collins & Stevens, 1982; Evens, Spitkowsky, Boyle, Michael, & Rovick, 1993; Gadd, 1995; Katz & O'Donnell, 1999; Leinhardt & Greeno, 1986). When a student makes an error during a problem-solving task, the tutor must decide: (a) whether to intervene, (b) if so, how—for example, what general approach should the tutor take—didactic or guided? (c) what tutoring tactics to use to carry out this approach, and (d) how deeply through the sub-topic hierarchy to take the discussion. The opportunity for post-solution reflection adds another variable to the instructional planning equation: the staging of instruction. At one extreme, the tutor can intervene immediately and address an error fully in its local context. At the other extreme, the tutor can ignore the error during problem solving and address it only during the post-solution dialogue.

Tutors' comments on the post-participation questionnaire suggest that their approach to staging instructional discussions fell at both extremes and various points in between. There appears to be three types of tutors, as illustrated by the comments in Table 7: *opportunists*, *stashers*, and *parcelers*. Opportunists tend to instruct in situ. They use the errors that students make during problem solving as opportunities for instruction. They might use the post-solution discussion to recap important points that came up during problem solving, but they rarely bring up new topics. (See, for example, Table 7,

[10] In partial correlations controlling for pretest score, overall gain was significantly correlated with numbers of abstraction dialogues ($p < .05$) and approached significance for number of tutor-initiated abstraction dialogues ($p = .06$), number of elaboration dialogues ($p = .13$), and a number of old dialogues ($p = .14$). In partial correlations with quantitative gain controlling for quantitative pretest, only the correlation with number of abstraction dialogues approached significance ($p = .12$)

comment 1). At the other extreme, stashers like to "squirrel away" things to talk about. They tend to ignore problem-solving errors, in the hope that students will recognize and correct them on their own. If a student doesn't catch a mistake, the tutor might hint at the error but generally tries not to intervene unless the student is stuck. Stashers mainly use post-solution discussions to reify the physical principle(s) that the student should have learned from his struggles. (See, for example, Table 7, comment 5.)

Table 7. Representative Comments From Tutors on the Staging of Instruction

From an "opportunist"
1. Generally, … I explained things as we went along with the problem. The only types of things that I usually explained after the problems was re-clarifying what I had said earlier to make sure that the student had it down for further usage.
From "parcelers"
2. There were times when I definitely took advantage of being able to explain things further after a problem, and then there were times when I had absolutely nothing to say after the problem was over.
3. I do not like to discuss the student's problem solving technique during the problem if I know that he is going to arrive at the correct answer doing it the student's way. I feel that this can disrupt their thinking process.
4. For the most part, I like to bring up relevant points or indicate mistakes in reasoning as they occur. However, if I knew that I had a "debrief" period, I might not insist upon discussing it right as the student made the error.
From a "stasher"
5. My strategy was often to let him go on doing the problem to let him get an answer. If he got a step wrong then I'd make a note of that and wait for the debrief. Then I'd let him search for his mistake and lead him to it giving clues on the way. If he got it right, we'd discuss the problem in perspective as to how to solve an entire class of similar problems, by exploring what variations could be made to the just-solved problem. Else, some technique or principle was clarified.

In-between these extremes are the parcelers, who divide instruction between problem solving and post-solution reflection in two ways: (a) through elaboration—that is, they partially address an error or topic during problem solving and then expand upon it during the post-solution dialogue (captured by comment 2 in Table 7) and (b) by addressing some errors or types of errors during problem solving, others afterwards (captured by comments 3 and 4 in Table 7).

The seven tutors' self-reports on their strategy suggested that three were opportunists, three were parcelers, and one was a stasher. However, cursory analyses of the problem-solving dialogues suggest that even the stasher intervened immediately more often than he might have thought; opportunists sometimes used the chance to discuss the problem after it had been solved more generously than they claimed; and parcelers appeared to sway to one extreme or the other under varying circumstances. For example, if the student was having trouble during problem solving or time was running short, the tutor

offered more direct advice during problem solving and kept the post-solution conversation short. If the student-tutor relationship was friendly, the tutor was more likely to reinforce problem-solving discussions during the reflective phase than if the relationship was strained. Future research should investigate the factors that constrain the staging of reflective instructional dialogues.

A Framework for Specifying Distributed Plans for Reflection

As shown in Figure 3, most reflective dialogues in the physics tutoring corpus elaborated upon problem-solving dialogues. If we consider only dialogues that dealt with domain content, 78 (48%) were distributed across problem solving and post-practice reflection (old—*elaborated* or *restated*), while 43 (26%) were coded as *new—domain*. The high frequency of distributed dialogues, coupled with the correlation between elaborated dialogues (and *old* dialogues in general) and learning, prompted us to examine the sub-corpus of tutor-initiated distributed dialogues more closely, in order to develop a framework for specifying distributed plans for reflection that can be implemented in ITSs.

Table 8 contains a sample distributed reflective dialogue. It illustrates one form of parceling described in the preceding section—addressing an error partially during problem solving, partially during the post-solution discussion. The distributed tutoring strategy can be informally described as follows:

> *During problem solving, flag the error or misconception; then let the process of solving the problem show the student that he or she was wrong. During post-solution reflection, make sure that the student recognizes that the solution contradicted his or her reasoning. Optionally, reify the general principle that the student missed.*

We dub this distributed strategy, "*Let the solution speak for itself."* In the problem-solving dialogue shown in Table 8, the student's first equation signals a common misconception: that all force problems deal with stationary objects.[11] This misconception is also evident in the student's reply to the tutor's question (problem-solving turn 2; "there's no acceleration"). The tutor flags the error by challenging the student's equation and pointing him to the problem statement (problem-solving turn 3; "…look at what the question is asking"). The student's remaining equations should bring out a contradiction in his reasoning, because they show that there *is* an acceleration in both the *x* and *y* directions, since there is a net force on the ring in both directions.

The tutor's main goal during the post-solution dialogue is apparently to ensure that the student realizes this contradiction. He first attempts to achieve this goal through a prompting query (post-solution turn 1; "So let me ask you a question; looking at your eqns. - is there any acceleration?"). When the student gives a partially correct response by acknowledging acceleration in the *x* direction (post-solution turn 2), the tutor prompts him to recognize the acceleration in the *y* direction (post-solution turns 3-6). Post-solution turn 7 is a striking example of conceptual generalization. The tutor reifies the principle that the student perhaps unwittingly applied in solving the current problem (NSL).

Like any dialogue plan, reflective dialogue plans have three essential elements: a main *goal* and its sub-goals, *tactics* for achieving these goals, and *content* to instantiate the speech acts (dialogue moves) that implement selected tactics. However, automated planning of *distributed* reflective dialogues requires some special considerations. For example, we need to consider how instructional goals should be parceled between problem solving and post-solution reflection, and how much of each goal should be achieved during each phase.

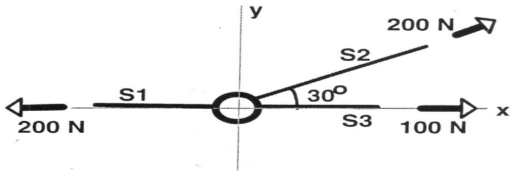
[11] It is possible that students acquire this misconception because force problems involving stationary objects are typically introduced in physics courses before force problems involving moving objects. Students come to assume that the net force equals zero.

With these considerations in mind, we propose the following framework for specifying distributed plans for reflection and illustrate it with reference to the dialogue in Table 8. The output of this framework is a *global plan* for reflection—an initial blueprint that needs to be modified dynamically in response to student input and other factors, such as the amount of time available in the tutoring session and the student's motivation level. Table 9 presents a global plan for the "*Let the solution speak for itself"* strategy implemented in the Table 8 dialogue. This plan is instantiated with the goals, sub-goals, tactics, and so forth applied in the distributed dialogue. Staging is indicated by the distribution of sub-goals between problem solving and post-solution reflection. Content is indicated by reference to specific dialogue turns in Table 8.

Table 8. Example of a Distributed Reflective Dialogue

Problem Statement:

In the figure below, each of the three strings exerts a tension force on the ring as marked. Use the labels S1, S2 and S3 to refer to the three strings. Find the components of the net force acting on the ring.



Problem-Solving Dialogue	Post-Solution Dialogue
The student writes the following equations in the Andes equations window: (1) Fnet=0 *The tutor interrupts to flag an error: the student wrongly assumed that there is no acceleration, so he set the net force (Fnet) equal to 0.* 1. T: Hold on; How do you know that Fnet = 0? 2. S: there's no acceleration 3. T: That's not for sure; maybe there _is_ acceleration; look at what the question is asking... *No more discussion occurs during problem solving. The student enters the following main equations (intermediate steps deleted) and solves the problem:* (2) Fnet_x= -S1+S3+S2cos(30) (3) Fnet_x = -200 + 100 + 200cos(30) = 73.2 (4) Fnet_y = 200sin(30) = 100N	1. T: So let me ask you a question; looking at your eqns. - is there any acceleration? 2. S: I guess so, is it because there is a force in the positive x direction. 3. T: Not only a force in the positive x-direction...are there any other net forces? 4. S: no 5. T: really? what is equation 4, then? 6. S: oh, the y direction also. 7. T: Yes. So you can see that there is a net force in the x-direction, which would result in an acceleration in the x-direction: Fnet_x = m * ax and similarly in the y-direction: Fnet_y = m * ay. This is a very important idea, and is called "Newton's Second Law". Is it familiar to you? 8. S: yes. 9. T: Good. Are you ready to move on, or do you have any more questions? 10. S: no. I am ready to move on. 11. T: Ok.

Proposed Framework for Specifying Distributed Plans for Reflection

Overall Goal*:* Goal specification involves two main elements:

1. *Goal Descriptor:* the type of goal, according to theories of intention in discourse (e.g., Mann & Thompson, 1988)—for example, *enabling* understanding or the ability to perform an action; *convincing* the hearer to carry out an action; *motivating* the hearer, and so forth.
2. *Object:* the knowledge piece that the tutor wants the student to understand or be motivated to learn; the action the tutor is trying to convince the student to take, and so forth. For example, the tutor's main goal in the dialogue shown in Table 8 is to enable the student to understand that force problems do not always involve stationary objects and NSL (a net force implies acceleration) applies to static and dynamic situations (Table 9, top).

Manner: How should goals and sub-goals be achieved? There are three main considerations:

1. *Staging:* Which goals or sub-goals should be achieved during problem solving, and which during the post-solution discussion? During problem solving in the example shown in Table 8, the tutor tries to make the student aware that he should not assume zero acceleration (Table 9, problem-solving Sub-goal-1). During the post-solution dialogue, the tutor accomplishes three sub-goals. He ensures that the student: (a) realizes that the solution contradicts his assumption of a zero net force (Table 9, post-solution Sub-goal-1), (b) understands the formalism for NSL (Table 9, post-solution Sub-goal-2), and (c) understands that his equations apply this principle (Table 9, post-solution Sub-goal-3).
2. *Scope of intervention:* To what extent should a goal or sub-goal be achieved, in whichever stage it occurs? What particular knowledge pieces should be addressed via a didactic, co-constructed explanation, and how deep should the explanation go? In the example, the tutor only hints at the student's misconception during problem solving. But the questions he asks strongly imply that acceleration is not zero, as the student assumed (Table 9, problem-solving Sub-goal-1). During the post-solution discussion, the tutor builds on this lesson by helping the student see *why* acceleration is not zero—because the net force in either direction does not equal zero (Table 9, post-solution Sub-goal-1). He then abstracts from the current situation, reifying the general principle that it illustrates, Newton's Second Law (Table 9, post-solution Sub-goals 2 and 3).
3. *Tactics:* What tutoring tactics—for example, hints, Socratic-style dialogues (a.k.a. "directed lines of reasoning;" Hume et al., 1996), didactic explanations—should be used to achieve each goal and sub-goal? For example, during problem solving, the tutor could have told the student directly that acceleration is not necessarily zero. Instead, he took a Socratic approach, first via a challenge query (Table 8, problem-solving turn 1, "How do you know that acceleration is zero?"), then with a hint (Table 8, problem-solving turn 3, "Look at what the question is asking"), and finally by letting the solution bring out the contradiction in the student's reasoning (see also Table 9, problem-solving Sub-goal-1, Tactics). Tutors frequently combine tactics like this. During the post-solution dialogue, the tutor first steers the student via questioning towards recognizing that the solution contradicted his reasoning (Table 9, post-solution Sub-goal-1, Tactics), and then shifts to a direct statement of NSL (Table 9, post-solution Sub-goals 2 and 3, Tactics).

Table 9. A Distributed Plan for Achieving the Main Goal of the Dialogue Shown in Table 8

<table>
<tr><td colspan="2">Overall Goal:
Descriptor: enable-understanding
Object: net force is not necessarily zero; there may be an acceleration in accordance with Newton's Second Law, although acceleration is not visually apparent
Constraints: evidence of misconception during problem solving</td></tr>
<tr><td>Problem-Solving Plan</td><td>Post-Solution Reflection Plan</td></tr>
<tr><td>Sub-goal-1 Descriptor: raise awareness
Object: net force is not necessarily zero
Constraints: evidence of misconception

Scope of intervention: flag misconception; there may be an acceleration
Constraints: there is acceleration in current situation

Tactics:

(1) hint (turns 1 and 3)
Constraints: Student understands that a net force of zero implies no acceleration (NSL).

(2) let solution contradict student's assumption (equations 2-4)
Constraints: student is able to carry out correct solution; otherwise, use repair dialogue(s) to help student generate correct solution</td><td>Sub-goal-1 Descriptor: enable-understanding
Object: net force is not necessarily zero
Constraints: evidence of misconception
Scope of intervention: verify misconception: since net force exists in x and y directions, there must be acceleration in both directions
Constraints: student understands that net force implies acceleration (NSL); otherwise, initiate repair dialogue to teach this relation
Tactics: prompting query (turn 1) OR directed line of reasoning (turns 1-6)
Constraints: if student response to prompt is incorrect, use directed line of reasoning until student recognizes acceleration in both directions
…….
Sub-goal-2 Descriptor: enable-understanding
Object: formalism for NSL in two dimensions
Constraints: student may not be able to relate his equations for net force to the formalism for NSL (F = ma)
Scope of intervention: teach general formalism for net force in two dimensions
Constraints: student may not be familiar with formalism for NSL
Tactics: didactic explanation (turn 7)
Constraints: None
…….
Sub-goal-3 Descriptor: enable-understanding
Object: current situation illustrates NSL
Constraints: student is familiar with NSL, but may not be able to recognize how it applies to a given situation
Scope of intervention: reify NSL
Constraints: student is familiar with NSL; repair dialogue otherwise
Tactics: didactic explanation
Constraints: query to check familiarity with NSL (turn 7)</td></tr>
</table>

Content: What speech acts should be used to implement tutoring tactics, and what subject matter should instantiate these speech acts? For example, the tutor's first hint during problem solving is

implemented via a *challenge* query (Table 8, problem-solving turn 1), while his second hint is implemented via a *directive* to look at the problem statement (Table 8, problem-solving turn 3). The tutor's didactic explanation of NSL is implemented via an *assertion* of general formalisms for deriving the net force in both directions, followed by an *assertion* of the physical principle that these formalisms represent (Table 8, problem-solving turn 7).

As illustrated in Table 9, constraints govern the distribution of instruction between problem solving and post-solution reflection (*staging*), the extent to which a sub-goal is addressed in each stage (*scope of intervention)*, and the selection of tutoring tactics. From an implementation standpoint, constraints form the "left-hand" (conditional) side of dialogue planning rules. For example, the tutor can only "let the solution speak for itself" if the student has the ability to generate a solution—in this case, produce equations that calculate the net force in the x and y directions. When a cognitive prerequisite such as this is not met, a repair dialogue addressing the requisite knowledge should be posted to the planning agenda. In the sample dialogue, the tutor issued a second hint during problem solving (Table 8, problem-solving turn 3) when the student gave an incorrect response to his initial attempt to flag the student's misconception (problem-solving turn 1). During the post-solution dialogue, the tutor had to "repair" the student's belief that acceleration takes place (only) in the x direction (post-solution turns 3-6). Some constraints block execution of dialogue-planning actions. For example, during the post-solution dialogue, the tutor seemed primed to execute a repair sub-dialogue about NSL. However, he cancelled this plan when he verified that the student was already familiar with this principle.

In addition to satisfying cognitive constraints (requisite background knowledge), reflective dialogue planners should satisfy affective constraints (e.g., the student's level of frustration) and logistical constraints (e.g., the amount of time remaining in the student's session). This applies both to generating a global plan, such as the one shown in Table 9, and to dynamic re-planning. For example, we have observed many instances of tutors shifting from interactive to didactic tactics when they notice that the student is getting frustrated or when a directed line of reasoning is heading nowhere. Achieving this level of responsiveness in automated planners presents a challenge, especially for student modeling and natural-language processing technology. Further research into which constraints govern all aspects of reflective dialogue planning—goal selection, distribution, and execution—and how these constraints do so, is needed to guide this effort.

Motivated in part by research that shows the effectiveness of reflection *during* problem solving (e.g., Bielaczyc, Pirolli, & Brown, 1995; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi & VanLehn, 1991; VanLehn, Jones, & Chi, 1992), several ITS researchers have developed techniques for analyzing reflective dialogues between a student and a tutor, and have implemented prototype reflective dialogue planning systems (e.g., Akhras & Self, 2000; Goodman, Soller, Linton, & Gaimari, 1998; Pilkington & Mallen, 1996; Ravenscroft & Pilkington, 2000). Some of these approaches were designed to support belief revision and conceptual change—as the tutor in the Table 8 dialogue attempts to do—and to handle multi-exchange interactions similar to those in which human tutors and learners engage. As such, they show promise for planning post-solution and distributed reflective discussions.

Is there a repertoire of distributed strategies that tutors use to achieve particular instructional goals? Besides the "let the solution speak for itself" strategy, there are several others that we observed. For example, each of the strategies described below was implemented by two or more tutors:

Strategy name: *Suggest alternate strategy/justify initial strategy*
Overall goal: Enable student to achieve a problem-solving goal more efficiently.
During problem solving: Advise the student about a more efficient strategy to use.

During post-solution reflection: Discuss how the student could have achieved the sub-goal with his original strategy, in part to demonstrate that the suggested strategy is more efficient.

Strategy name: *Hint at schema/reify schema*
Overall goal: Enable the student to recognize a common problem schema.
During problem solving: If the tutor suspects at the start of the problem that the student will not be able to derive a global plan for solving it, hint at the problem schema (e.g., "Think of Newton's Second Law", "We did a problem like this last week").
During post-solution reflection: Reify the schema (e.g., "You see that all these problems have to do with work done by a variable force, right?").

Strategy name: *Challenge correct solution*
Overall goal: Check understanding of principles underlying correct actions.
During problem solving: If solution is error-free, no intervention.
During post-solution reflection: Verify understanding of principles associated with problem-solving actions. A common tactic for doing this is to pose "what if" scenarios (problem variations) that challenge the student to apply the same principle to a modified physical situation.

Although we identified several recurring distributed strategies in the physics tutoring corpus, this does not necessarily imply that tutors implement these strategies intentionally. We did, however, find evidence that tutors did some amount of deliberate planning during problem solving. For example, in a few cases, the tutor explicitly deferred a topic until the post-solution dialogue (e.g., "We can discuss this later on," or "I'll give you an example to show why your thinking is flawed after you solve this problem"). Also, in their answers to the survey question about whether the session format (debrief or no-debrief) influenced their behavior, four of the seven tutors claimed that it did. This is illustrated by the comments in Table 10. However, since self-reports can be unreliable (e.g., Nisbett & Wilson, 1977), we plan to investigate the relationship between session format and tutoring behavior in future analyses. These analyses will focus on specific speech acts. We expect, for example, that tutors who claimed to offer more direct guidance when debrief was suppressed (e.g., Table 10, comments 1-3) will tend to make more advisory assertions and ask fewer diagnostic questions during problem solving in the no-debrief sessions than in the debrief sessions. These analyses will further our understanding of the instructional roles of post-solution reflection and the ways that human tutors distribute instruction between the latter and problem solving.

Table 10. Tutors' Comments on How Problem Format Affected Their Behavior

1. On the occasions that we were not allowed to talk afterwards, I helped the student along more when he/she was stuck because their interest was in getting the answer….Also, I feel that I gave less direct help on the debrief questions and made the student work harder to get the answer.
2. When we were not permitted to discuss the problems I was, in my mind, more likely to help or prompt a question to the student to try and direct the students thinking process towards the correct way of solving the problem. When we were permitted to discuss I felt that I was more apt to letting the student make more mistakes without jumping in. Especially if I thought that it may be a learning opportunity for the student to make the mistake.
3. Primarily, I found myself to-the-point when there were no debrief sessions allowed.
4. My interactions with the student would differ a little bit if I knew that I could continue the discussion afterwards. For the most part, I like to bring up relevant points or indicate mistakes in reasoning as they occur. However, if I knew that I had a "debrief" period, I might not insist upon discussing it right as the student made the error.

A STUDY OF THE EFFECTIVENESS OF POST-SOLUTION REFLECTION QUESTIONS

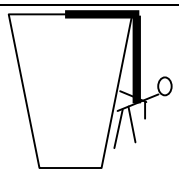
Consistent with our previous studies of post-solution dialogues (e.g., Katz et al., 2000), the first study discussed in this paper suggested that one of the primary roles of these dialogues is to teach the conceptual knowledge underlying strategic, problem-solving knowledge. It also showed that students who learned more (as measured by pre-test to post-test gain score) had more post-solution discussions with their tutor, especially discussions that abstracted from the current problem-solving situation. However, the data from this study could not tell us whether students who engage in post-solution reflection outperform students who do not. The primary goal of the second study was to address this question. In particular, it investigated the extent to which asking the kinds of reflection questions that tutors in the first study posed to students, coupled with feedback on students' responses, supports understanding of physical concepts and principles and transfer—students' ability to solve similar problems to those in Andes. A secondary goal of this study was to determine whether natural-language interaction with a human tutor during question answering is more effective than receiving canned text as feedback. Addressing this question provides insight into the importance of incorporating natural-language processing techniques in reflective dialogue planning modules.

To address these questions, we conducted an experiment with three conditions. In the first experimental condition, students received reflection questions in Andes that were followed by canned feedback. In the second experimental condition, students received the same set of reflection questions, but they were able to interact with a human tutor via teletype while answering them. (See Table 11 for an example of an Andes problem with its reflection questions.) In the control condition, students solved

problems in Andes but did not receive the reflection questions or feedback. Using a pre- and post-test design, we compared the reflection question groups to each other and to the control group, with respect to gain scores.

As we saw in the first study, tutors and their students focused on conceptual knowledge during the post-solution dialogues (See Table 2). For this reason, the reflection questions used in the second study focused on physics concepts and principles and on how to apply them to problem solving. Most of these questions were adapted from those that tutors asked in the first study. Other questions were written by physics instructors, one of whom participated in the first study. Each Andes problem contained three to eight reflection questions. All students in the experimental conditions received the same set of reflection questions per problem, in the same order. There was no attempt to tailor the reflection questions to students' individual needs. As such, this study provides an indication of the value of asking reflection questions and providing feedback on students' responses, and compares the merits of canned feedback with adaptive, human feedback. However, it gives no indication of the value of adaptive selection of reflection questions.

Table 11. Sample Problem and Reflection Questions

Problem: A rock climber of mass 55 kg slips while scaling a vertical face. Fortunately, her caribiner holds and she is left hanging at the bottom of her safety line. Find the tension in the safety line. 
Reflection Questions: Suppose Sir Isaac Newton came back to life just to help you solve this problem. As a hint, he told you to think of his 2nd law of motion. Use it to explain why the tension equals the weight in this particular problem. Suppose the maximum tension in the rope was 500 N. What would happen to the climber if she hung stationary on the rope? What minimum acceleration must the climber have in order for the rope not to break while she is rappelling down the cliff? (You do not have to come up with a numerical answer. Just solve for "a" without any substitution of numbers.) Suppose the climber were rappelling down the rope with a constant velocity equal to or less than the minimum acceleration found in the previous question. Would the rope still break?

Method

Participants were 46 paid volunteers recruited from introductory physics classes at the University of Pittsburgh. There were 15 students (6 male and 9 female) assigned to the canned feedback condition, 16 students (4 male, 12 female) to the human tutored feedback condition, and 15 students (9 male and 6 female) to the control condition. The canned feedback and control conditions were run prior to the human feedback conditions, and students were randomly assigned to one of these two conditions. The human tutored feedback condition was run in two phases, with a different physics expert playing the role of tutor in each phase. The human tutor interacted with students only during the reflection question phase, not while students were solving problems. Students in all three conditions used Andes' automated coaching during problem solving (see Figure 1).

Each student completed the study over several sessions. Mean completion time was 12.3 hours (*SD* = 2.7 hr). Students completed a background survey that solicited demographic information, prior physics

training, and Math and Verbal SAT scores. They also took a math test and the physics pre-test. The physics pre-test and post-test each covered the same topics and contained 36 questions: 9 quantitative mechanics problems similar to those that students worked on in Andes, and 27 qualitative questions that each tested a mechanics concept or principle. The order in which the two tests were administered was counterbalanced.

After the pre-test, students reviewed a workbook chapter on kinematics developed for the experiment. Students next received training in the use of Andes and worked on three kinematics problems. At this point the three conditions differed for the first time. In the canned feedback and human feedback conditions, reflection questions were presented after the student solved each problem and the correctness of the student's answer was confirmed.[12] For each reflection question, the student typed in a response and pressed "Enter." In the canned feedback condition, a canned response written by a physics expert was displayed after the student answered each reflection question. For example, the canned feedback provided on the first question in Table 11 ("…Use [Newton's 2nd law of motion] to explain why the tension equals the weight in this particular problem.") reads as follows:

> *There is no acceleration (in any direction) since the climber has a constant velocity (of zero). Newton would say since the acceleration in this problem is 0 then that means that there exists no net force. In other words, although forces may act on the object, the sum of all those forces must add up to 0. When we calculate all the forces, we see that there are two forces that act on the system: F_net = T - mg. Since the acceleration is 0, F_net is 0, so T = mg.*

In the human feedback condition, students could enter a response or a question to the human tutor and engage in a series of teletyped exchanges until the tutor was satisfied that the student was ready to move on to the next question. Students in the human tutored feedback condition experienced a delay in receiving their tutor's teletyped responses. To control for this factor, we implemented a delay in students' receipt of the canned responses in the canned feedback condition. The delay for receiving each canned response was set to 22 seconds—the average time required for a human to type these responses, at an estimated average typing speed of 40 words per minute (8 characters per second).[13] In the control condition, students simply went on to the next problem after solving the current problem.

After working on the three kinematics problems, students reviewed a workbook chapter on dynamics and solved three dynamics problems in Andes. Once again, students in the two experimental conditions went through the reflection questions in the manner described above; those in the control condition did not. In the remaining sessions, students solved additional kinematics and dynamics problems in Andes. The experimental groups did 6 more Andes problems—3 more kinematics problems and 3 more dynamics problems—yielding a total of 12 problems, each followed by reflection questions. To control for differences in time on task due to the reflection questions, the control group worked on 3 more problems than the experimental groups—6 more kinematics problems and 3 more dynamics problems—yielding a total of 15 problems. After completing the curriculum, students took the post-test, filled out a post-participation questionnaire, and were paid.

---

[12]Students eventually solve each Andes problem, because hinting is designed to be increasingly directive. If the student is lost and makes several successive help requests, he is eventually given an equation that, when instantiated correctly, will solve the problem.

[13]Initially, we calculated a separate delay for each canned response, using the estimated 8 characters/second typing rate. However, some of the canned responses are quite lengthy so the delays associated with these responses were lengthy too—in some cases, two minutes or more. During pilot testing, we noticed some students' impatience with the long delays and decided that such delays could potentially handicap students in the canned feedback condition. For this reason, we decided to use the *average* delay calculated for the canned responses as a uniform delay across reflection questions. The average delay was 22 seconds.

Results and Discussion

The primary measure of student learning was the overall gain score, defined as the difference between the percentage of questions correct on the post-test and pre-test. The 24 items on the pre-test and post-test that were not multiple-choice were scored by two independent raters. Initial agreement was 91% on the pre-test answers (kappa = .83) and 93% on the post-test answers (kappa = .82), and all disagreements were resolved by discussion before calculating gain scores. Student gain scores were entered into a one-way ANOVA with condition as a between-subjects factor[14]. Our primary prediction was that students would learn more with reflection questions and some form of feedback (the canned feedback and human tutored feedback conditions) than without the same (the control condition). This prediction was confirmed. The main effect of condition was significant, F (2, 43) = 7.2, $p < .01$. Pair-wise comparisons using the Tukey HSD test with an alpha level of .05 revealed that the mean gain scores for both the canned feedback condition (25.0%) and the human tutored feedback condition (19.5%) were significantly greater than the gain score for the control condition without reflection questions (8.0%). The difference between the canned feedback and tutored feedback conditions did not approach significance ($p > .4$).

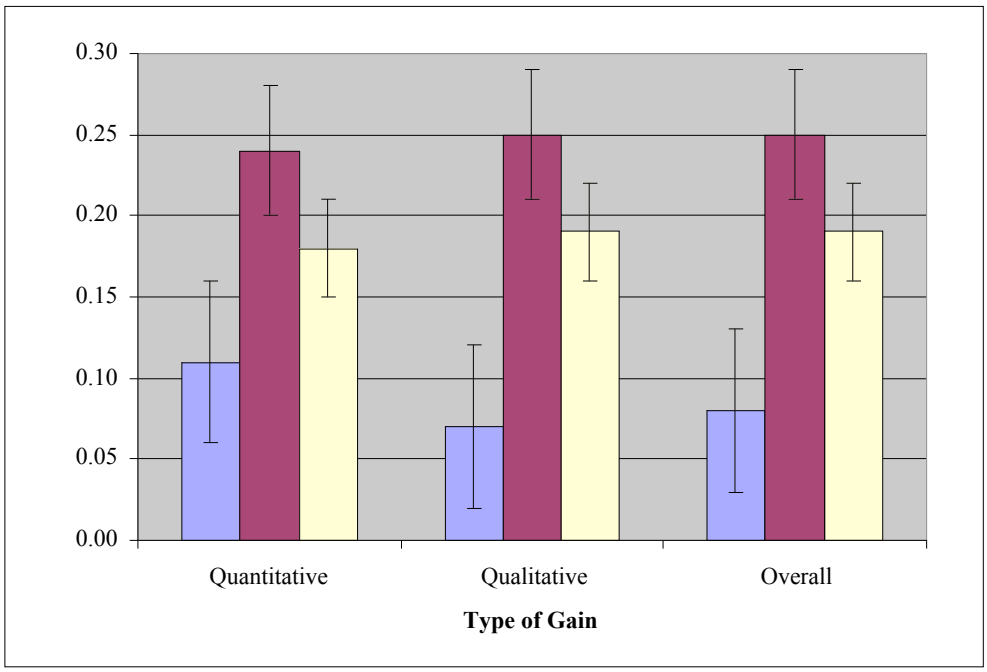


Figure 5. Mean gain scores from pre-test to post-test, by condition

[14] Many methodologists (e.g. Cook & Campbell, 1979) recommend against the use of gain scores as a dependent variable in statistical analyses. The alternative analysis generally adopted is an analysis of covariance approach, with post-test scores as the dependent variable and pre-test scores as a covariate (Cronbach & Furby, 1970). Allison (1990) and Maris (1998) have shown, however, that in some circumstances analyzing gain scores is preferable to using the covariate approach. Furthermore, an analysis of gain scores with pre-test scores as a covariate is formally equivalent to a covariate approach using post-test scores as the dependent measure (Werts & Linn, 1970). For each of the ANOVA and ANCOVA analyses of gain scores reported in this section, re-analysis using the covariate approach (with post-test as the dependent variable) produced identical patterns of results and statistical significance. For clarity of exposition, we have chosen to present the analyses based on gain scores rather than those based on covariate-adjusted post-test scores.

In order to assess whether reflection questions affected quantitative problem-solving skills and conceptual understanding differently, gain scores were also calculated separately for the 9 quantitative problems and the 27 qualitative questions. The mean quantitative, qualitative, and overall gain score for each condition (reflection-question vs. control) is shown in Figure 5 along with standard error bars. The quantitative and qualitative gain score analyses revealed the same pattern of results as the overall gain scores, with greater gain scores in both reflection-question conditions than in the control condition. The differences were statistically significant for the qualitative gain scores, F (2, 43) = 6.6, $p < .01$, but not for the quantitative gain scores, F (2, 43) = 1.6, $p = .21$, probably because there were fewer quantitative problems per participant, resulting in greater variability in the quantitative gain scores. Thus, reflection questions with feedback produced greater learning than problem solving alone on all three measures of gain, although not significantly so on quantitative gain scores.

The analysis of the primary measure of learning (overall gain scores) was repeated, adding five covariates to the model. The covariates were selected from our measures of background, aptitude, and time on task based on preliminary analyses of whether these factors differed among the three conditions. Significant differences among the three conditions were found for two of the measures on the background survey: which of two physics courses participants were currently enrolled in—algebra-based or calculus-based—($\chi^2(2) = 9.5$, $p < .01$), and amount of high school physics background (F (2, 42) = 3.9, MSe = 0.32, $p < .05$). A greater proportion of students in the control condition (80%) were currently enrolled in the algebra course compared to students in the canned feedback (47%) and human feedback (25%) conditions. As for high-school physics background, more students in the canned feedback condition (40%) had not taken any physics course in high school compared to students in the human feedback (6%) and control (7%) conditions. Both of these factors were therefore included as covariates. Significant differences among the conditions were also found for two measures related to time on task: the number of days the participant took to complete all of the sessions, F (2, 43) = 5.8, MSe = 57.9, $p < .01$, and the total number of hours each participant spent in all of the sessions combined, F (2, 43) = 5.3, MSe = 6.1, $p < .01$. Participants took an average of 11.0 hours to complete the sessions in the control condition, 12.0 in the canned feedback condition, and 13.8 in the human feedback condition, although only the difference between the human feedback condition versus the other two was significant by a Tukey HSD test. In addition to the four measures that differed among the conditions, we also included the physics pre-test scores as a fifth covariate even though there were no significant differences among conditions on the pre-test, F (2, 43) = 0.9, MSe = .03, *ns*. The mean pre-test scores were .61, .53, and .54 (SE = .04) for the control, canned feedback, and human tutored feedback conditions, respectively.

The re-analysis including these five covariates confirmed our initial findings. Controlling for these five factors, the effect of condition on overall gain scores remained significant, F (2, 37) = 3.3, $p < .05$. The increase in learning that resulted from the inclusion of reflection questions therefore does not appear to be an artifact of differences in background knowledge, aptitude, nor time on task.

We also examined the number of Andes coaching hints received during problem solving in each condition. As addressed earlier in the overview section on Andes, students could access on-demand hints while working with Andes.[15] There was a marginal difference between conditions on mean number of hints received, with 236.1 in the canned group, 289.4 in the human tutored group, and 413.7 in the control group (F (2, 43) = 2.64, $p = .08$). Note that this pattern of differences runs counter to the pattern of results for gain scores (see Figure 5), such that students with higher gain scores tended to ask for fewer

[15] Additionally, during some student sessions Andes provided a handful of unsolicited interface warning messages that also contained content similar to those of the solicited hints (see also Footnote 3). These warning messages were therefore counted along with the solicited hints; however, they constituted only 1% of the total hint count (188 warnings vs. 14190 solicited hints across all 46 sessions).

hints, and vice versa.[16] Although not statistically significant, these differences in hint counts suggest that control students who received no reflection feedback felt a greater need for additional help during their sessions, even if the additional help did not translate to better gain scores.

Having established that reflection questions with feedback led to increased learning, we have begun to investigate the nature of this effect further. Our first step has been to examine the relationship between gain scores and various background and performance measures within the reflection question conditions. For the 10 students in the canned feedback condition for whom self-reported SAT scores were available, SAT Math, Verbal, and total scores were all positively correlated with overall gain score (r = .50, .48, and .63 respectively), although only the correlation for total SAT was significant ($p < .05$). None of the SAT measures were correlated with gain for the students in the human feedback condition, however; r = -.07, -.09, and .02 for the Math (N = 11), Verbal (N = 11), and total (N = 14) SAT score respectively.

For all 15 students in the canned feedback condition, we also examined the relationship between overall gain score and two performance measures: average time spent answering reflection questions (RQ-time) and average number of words (RQ-words) produced in answers to reflection questions. Both of these measures were correlated with overall gain, although only the latter was significantly so (r = .61, $p < .05$); time was only marginally so (r = .39, p =.15). Thus, the more students said in response to reflection questions the more they learned, as measured by the difference between pre-test and post-test scores. This suggests that self-explanation may at least partially account for the effect we have observed (e.g., Chi et al., 1989). To test this and other hypotheses, we will focus future analyses on the content of students' responses. For example, we will determine whether gain scores were correlated with the number of principles that students expressed, the number of relationships between principles that they expressed, whether or not they answered questions correctly, and so forth.

Similar correlational analyses will be conducted on the human-tutored data, with the goal of identifying features of students' and tutors' interaction about the questions that promoted (and failed to promote) learning. We expect that these analyses will not only offer an explanation as to why there was, surprisingly, no significant difference between the two experimental conditions, but also some insight into how to design effective reflective dialogue planners for ITSs. Although the second study calls into question the need for natural-language processing in reflective modules, it does so only with respect to the learning conditions carried out in this study—that is, reflective dialogues about a pre-specified set of questions. Whether reflective discussions that are tailored to students' individual needs are more effective with natural language than with canned feedback is a question for further research.

More research is also needed to tease apart the effectiveness of the reflection questions from that of the feedback (canned or human-provided). This study showed that reflection questions, with either form of feedback, support learning. But it does not allow us to assign credit to either intervention—the reflection questions or the feedback—or to both interventions in combination. A follow-up study, with additional conditions (e.g., reflection question with no feedback, canned feedback only—no reflection questions), is needed to address this question.

GENERAL DISCUSSION AND CONCLUSIONS

The field of education is rife with "good practices" that we know very little about. Sometimes we know that a "good practice" is effective, but understand little about how or why it works. One-on-one human tutoring is a case in point, although several studies in recent years have made significant strides in uncovering some of the reasons for the observed two-sigma effect of tutoring over classroom instruction (Bloom, 1984)—for example, co-construction of knowledge (Chi, Siler, Jeong, Yamauchi, & Hausmann,

[16] Hint counts were uncorrelated with gain scores (r = -.04) and therefore were not used as a covariate.

2001; VanLehn, Siler, Murray, Yamauchi, & Baggett, in press). Sometimes we neither know whether a "good practice" really works nor, if it does, why. Debriefing—reflective discussions of practice exercises—is an example of this.

The two studies described in this paper take a step towards answering these questions about post-solution, reflective dialogues. The first study identified several instructional roles of these dialogues in physics tutoring, showed that they are tailored to students' individual learning needs, demonstrated their effectiveness—especially when they are plentiful and generalize from the current problem—and uncovered several distributed strategies for reflection. The second study showed that one type of reflective activity—namely, reflection questions with live or canned feedback—can support learning. To our knowledge, this is the first controlled experiment about post-solution conversations that demonstrates their instructional value.

We conclude with some remarks about the potential of post-solution discussions to facilitate "caring for learners." In our discussion of the first study, we pointed out some ways that adaptive instruction took place during post-solution reflection. The tutors focused on different types of knowledge to varying degrees with different students—for example, they spent more time on strategic knowledge with weaker students than with students who had more background knowledge. They also elaborated on the concepts and strategies discussed during problem solving with weaker students. With more able students, the tutors tended to discuss procedural issues, such as more efficient ways of achieving a particular problem-solving goal, and the conceptual basis for problem-solving actions ("why did you do what you did?").

When we consider the various information types that we have discussed more closely—that is, restatement, elaboration, and bringing new topics to the table—we catch a glimpse of some more subtle ways that post-solution dialogues can support adaptive instruction and possibly enhance the instruction provided during problem solving. The information status categories account for various ways that post-solution reflection helps tutors adhere to Grice's (1975) conversational maxims. Restating points made during problem solving, possibly in clearer terms, can enhance the *manner* of instruction. Elaborating on lessons begun during problem solving supports the maxim of *quantity*—not saying more than is needed in a particular context. Post-solution, elaborative discussions may reduce the learner's cognitive load during problem solving and, possibly, the tutor's load as well. Deferring some topics to post-solution reflection can also achieve this, while supporting the maxim of *relevance.* It would be quite odd, for example, for a tutor to assess a student's general ability in physics during problem solving.

Perhaps the main way that post-solution reflection can support tutors' efforts to care for learners is by enhancing student performance. Both studies showed this, and the second study suggested that even non-adaptive reflective dialogues enhance learning. What accounts for the effect we observed—was it the question-answering activity itself, or the fact that it took place after the problem had been solved? In other words, what is the most important aspect of post-solution reflection: the *post-* part, or the *reflection* part? Would we have observed the same effect (or an even better one) if we had asked the same questions *during* problem solving? This is tantamount to asking: Which staging strategy is best: that of opportunists, stashers, or parcelers? Perhaps opportunists, who conduct learning conversations almost entirely during problem solving, have the right idea. They exploit all the benefits of contextualized, "just-in-time" learning. But this comes at a potential cost: Too much information during problem solving could be distracting and might increase the student's cognitive load (e.g., Sweller, 1988). Stashers shield their students from excessive interruptions by offloading most of the instruction until post-solution reflection. But this might make contextualizing lessons difficult. Perhaps parcelers strike a happy balance between these two extremes, by elaborating on problem-solving lessons during the reflective phase, and by discussing some problems on the spot and deferring others. If parceling is the most effective approach, we are left with the question of which topics, or which *aspects* of topics, to address when. Further research on these issues will allow us to better understand and optimize the "good practice" of debrief.

Acknowledgments

This research was supported by grants from the Spencer Foundation (grant number 199900054) and the Office of Naval Research, Cognitive Science Division (grant number N00014-97-1-0848). The data presented and views expressed are not necessarily endorsed by these agencies. James Carlino, Beth Nicholson, Annalia Palumbo, and Paul Reilly contributed to the research. Kurt VanLehn and the anonymous reviewers provided many helpful and interesting comments on an earlier version of the manuscript.

References

Akhras, F. N., & Self, J. A. (2000). System intelligence in constructivist learning. International Journal of Artificial Intelligence in Education, 11, 344-376.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20*, 93-114.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13(2),* 221-252.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Bransford, J. D., Sherwood, R. S., Hasselbring, T. S., Kinzer, C. K., & Williams, S. M. (1990). Anchored instruction: Why we need it and how technology can help. In D. Nix & R. Spiro (Eds.), *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 115-141). Hillsdale, NJ: Erlbaum.

Brown, J. S. (1985). Process versus product: A perspective on tools for communal and informal electronic learning. *Journal of Educational Computing Research, 1,* 179-201.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471-533.

Chi, M. T. H., & VanLehn, K. (1991). The content of physics self-explanations. *Journal of the Learning Sciences, 1*, 69-105.

Collins, A., & Stevens, A. L. (1982). Goals and strategies of inquiry teachers. In Glaser, R. (Ed.), *Advances in instructional psychology*, vol. 2 (pp. 65-119). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cronbach, L. J., & Furby, L. (1970). How should we measure change – Or should we? *Psychological Bulletin, 74*, 32-49.

Derry, S., & Lesgold, A. (1996). Toward a situated social practice model for instructional design. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 787-807). New York: Macmillan.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston: Heath.

Evens, M. W., Spitkowsky, J., Boyle, P., Michael, J. A., & Rovick, A. A. (1993). Synthesizing tutorial dialogues. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 137-142)*, Boulder, CO.* Hillsdale, NJ: Erlbaum.

Gadd, C. S. (1995, July). A theory of the multiple roles of diagnosis in collaborative problem solving discourse. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 352-357), Pittsburgh, PA.

Gertner, A., & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. In *Proceedings of the 5th International Conference, ITS 2000*, Montreal, Canada.

Goodman, B., Soller, A., Linton, F., & Gaimari, R. (1998). Encouraging student reflection and articulation using a learning companion. *International Journal of Artificial Intelligence in Education, 9*, 237-255.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359-387.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics*, vol. 3 (pp. 41-58). New York: Academic Press.

Hume, G. D., Michael, J., Rovick, A., & Evens, M. (1996). Hinting as a tactic in one-on-one tutoring. *Journal of the Learning Sciences*, *5(1)*, 23-47.

Katz, S., & Allbritton, D. (2002). *Improving learning from practice problems through reflection*. Paper presented at the annual meeting of the *American Educational Research Association*, New Orleans.

Katz, S., & O'Donnell, G. (1999). The cognitive skill of coaching collaboration. In C. Hoadley & J. Roschelle (Eds.), *Proceedings of Computer Support for Collaborative Learning (CSCL) 1999* (pp. 291-299), Stanford, CA.

Katz, S., O'Donnell, G., & Kay, H. (2000). An approach to analyzing the role and structure of reflective dialogue. *International Journal of Artificial Intelligence and Education, 11*, 320-343.

Katz, S., Lesgold, A., Hughes, E., Peters, D., Eggan, G., Gordin, M., & Greenberg., L. (1998). Sherlock 2: An intelligent tutoring system built upon the *LRDC Tutor Framework*. In C. P. Bloom & R. B. Loftin (Eds.), *Facilitating the development and use of interactive learning environments* (pp. 227-258). Mahwah, NJ: Erlbaum.

Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, *4*(3), 187-210.

Leinhardt, G., & Greeno, J. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, *78*(2), 75-95.

Lesgold, A., Katz, S., Greenberg, L., Hughes, E., & Eggan, G. (1992). Extensions of intelligent tutoring paradigms to support collaborative learning. In S. Dijkstra, H. Krammer, J. van Merrienboer (Eds.), *Instructional models in computer-based learning environments* (pp. 291-311). Berlin: Springer-Verlag, 291-311.

Mann, W., & Thompson, S. (1988). Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, *8(3),* 243-281.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3*(3), 309-327.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction, 7*, 197-244.

Moore, J. D. (1996). Making computer tutors more like humans. *Journal of Artificial Intelligence in Education*, *7(2),* 181-214.

Moore, J. D., Lemaire, B., & Rosenblum, J. A. (1996). Discourse generation for instructional applications: Identifying and exploiting relevant prior explanations. *Journal of the Learning Sciences, 5(1),* 49-94.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Person, N. K., Graesser, A. C., Kreuz, R. J., Pomeroy, V., & the Tutoring Research Group. (2001). Simulating human tutor dialog moves in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 23-39.

Piaget, J. (1976). *The grasp of consciousness: Action and concept in the young child.* Cambridge, MA: Harvard University Press.

Pilkington, R.M., & Mallen, C. (1996). Dialogue games to support reasoning and reflection in diagnostic tasks. *Proceedings of the European Conference on Artificial Intelligence and Education* (pp. 220-225), Lisbon, Portugal.

Pioch, N. J., Roberts, B., & Zeltzer, D. (1997). A virtual environment for learning to pilot remotely operated vehicles. *Proceedings of Virtual Systems and Multimedia 1997*, Geneva, Switzerland.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal*, *24,* 13-48.

Ravenscroft, A., & Pilkington, R. M. (2000). Investigation by design: Developing dialogue models to support reasoning and conceptual change. *International Journal of Artificial Intelligence in Education, 11,* 273-298.

Rosé, C. P. (1997). The role of natural language interaction in electronics troubleshooting. *Proceedings of the Eighth Annual International Energy Week Conference and Exhibition*.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., & McPherson, J. A. (1998). Team dimensional training: A strategy for guided team self-correction. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 271-297). Washington, DC: APA.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning: *Cognitive Science*, *12*, 257-285.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences, 2,* 1-59.

VanLehn, K., Siler, S., Murray, C, Yamauchi, T. & Baggett, W. B. (in press). Human tutoring: Why do only some events cause learning? *Cognition and Instruction*.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Los Altos, CA: Morgan Kaufmann.

Werts, C. E., & Linn, R. L. (1970). A general linear model for studying growth. *Psychological Bulletin, 73*, 17-22.

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3-117.

White, B., Shimoda, T.A., & Frederiksen, J. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education, 10*, 151-182.