When is Reading Just as Effective as One-on-One Interactive Human Tutoring?

Kurt VanLehn (vanlehn@cs.pitt.edu), Art Graesser (agraesser@memphis.edu), G. Tanner Jackson (gtjacksn@memphis.edu), Pamela Jordan (pjordan@pitt.edu), Andrew Olney (aolney@memphis.edu), Carolyn P. Rosé (cprose@cs.cmu.edu)

Learning Research and Development Center, University of Pittsburgh,
Pittsburgh, PA 15260 USA
and
Institute for Intelligent Systems, University of Memphis,
Memphis, TN USA

Abstract

Many human tutorial dialogues have a similar format (Graesser, Person & Magliano, 1995) wherein the tutor poses a question, the student answers it, and they work collaboratively to improve the student's answer. We hypothesized that the well-known effectiveness of human tutoring was due to the last phase, and in particular, to the interactive nature of the tutorial dialogue there. To test the hypothesis, we compared tutoring which held constant the first two steps, the question and the student's initial answer, and varied whether the student engaged in tutorial dialogue or simply read an explanation of how to derive a correct answer. To our surprise, the initial experiment showed equal learning gains. Two further experiments painted a confusing picture, but it is safe to say that interactive tutoring is *not* always superior to reading.

Keywords: Natural language tutoring, dialogue, tutoring, learning, qualitative physics

Introduction

In the scientific literature, the educational enterprise, and the general public, expert human tutors are believed to provide better instruction than computer tutors or classroom teaching. Many studies (reviewed below) have tried to isolate the source of human tutors' effectiveness. The present paper adds to the accumulating evidence.

First, let us describe the type of tutoring under study. Graesser, Person and Magliano (1995) found that much human tutoring followed a 5-step pattern:

1. The tutor poses a question or problem.
2. The student attempts to answer it.
3. The tutor provides brief evaluative feedback.
4. The tutor and student collaboratively improve the answer or solution. This can involve a moderately long dialogue.
5. The tutor ends the discussion, often by asking the student if they understand, and almost always getting a positive response.

Here is an example:

1. *Tutor*: What does a t-test tell you?
2. *Student*: It tests whether a mean is significant.
3. *Tutor*: Sorta.
4. *Tutor*: Can it be applied to experiments with just one group, or do you need two or more groups?
   Student: More than one.
   Tutor: Right. Because the t-test compares the means of the two groups. What does it tell you about the two means?
   Student: Whether they are significant.
   Tutor: Almost. What you care about is whether one mean is really and truly higher than the other, or whether the difference between them is just an accident of sampling. Does the t-test tell you anything about that?
   Student: Yes.
   <etc.>
5. *Tutor*: So do you understand the t-test now?
   Student: Yes.
   Tutor: Good. Let's go on.

Graesser et al. (1995) hypothesized that the effectiveness of tutoring lies in the tutorial dialogue of step 4, and in particular, that it is the interactive nature of that dialogue which accelerates learning. For instance, learning would probably be harmed if students simply listened or read an explanation at step 4 instead of participating in a dialogue.

This hypothesis is consistent with several earlier studies of human tutoring. Wood, Wood and Middleton (1978) had a human tutor implement 4 different strategies for teaching preschool children how to assemble a complicated block structure. One strategy implemented the following rule: "If the child succeeds, when next intervening offer less help. If the child fails, when next intervening take over more control." (Wood et al, 1978, pg 133). The other strategies were less interactive. For instance, the least interactive strategy had the tutor just demonstrate the to-be-learned procedure. As predicted by the interaction hypothesis, the most interactive tutoring strategy produced the best performance on a post-test.

Swanson (1992) compared the highly interactive tutoring strategy of Wood et al. (1978) to simply lecturing. As in the Wood study, the same tutor implemented both forms of instruction, but Swanson's students were college students learning how lens work. As predicted by the interaction hypothesis, the more interactive tutoring produced more gains. Swanson also found that a second tutor could not learn to be interactive, and tended to lecture in both conditions.

Chi et al. (2001) took advantage of the propensity of untrained tutors to lecture, and first had a group of tutors work with tutees naturally. These tutors were then trained to be more interactive, e.g., by using content-free prompting as much as possible. Unlike Swanson's study, this training succeeded; analyses of the natural and trained dialogues showed that tutors did most of the talking when untrained, and students did most of the talking after the tutors were trained. Contrary to the interaction hypothesis, the learning gains of tutees in the two groups did not differ. However, Chi et al. did find the same correlation that Wood et al. found, which is that students who learned more also were more interactive during tutoring.

Rosé, Moore, VanLehn and Allbritton (2001) compared Socratic and Didactic strategies for tutoring students on basic electricity. Post-hoc analysis of the transcripts indicated that Socratic tutoring was indeed more interactive than the Didactic tutoring. However, the learning gains of the Socratically tutored students were not reliably different from those of the Didactically tutored students, although there was a trend in the expected direction.

Katz, Connelly and Allbritton (2003) compared interactive human tutoring from trained tutors to simply reading a text. In particular, they had a computer present a question (step 1 of the 5-step frame), and the student type in a paragraph-long answer (step 2). Students in the reading condition would then read a paragraph-long version of the correct answer. In contrast, students in the human tutoring condition had a computer-mediated (typed) dialogue with an expert human tutor (step 4). Although the tutorial dialogue showed little lecturing, the tutored students did not learn more than the reading students, contrary to the interaction hypothesis.

All the studies discussed so far have used human tutors, so it is difficult to establish that they cover the same material with all students. Using computer-based natural language tutors, such control is easier to obtain because the content of the dialogue is designed into the tutors. Rosé et al. (2003) compared computer-based natural language tutoring to reading multi-paragraph explanations that were written to have the same content as a maximally long tutorial dialogue. They found that tutored students learned no more than students who read the content instead of interacting with the computer tutor.

Graesser et al. (2001) compared reading a computer-literacy textbook to natural language computer tutoring that was designed specifically to emulate the tutorial dialogues found during step 4 of the 5-step frame. As predicted by the interaction hypothesis, the tutored students learned more than the students who studied the textbook. This result was also found with our task domain, qualitative physics (Graesser et al., 2003). However, in both these studies, students who studied the textbook did not answer questions. That is, not only was step 4 missing, so were the other steps as well. So Jackson et al. (2004) repeated the physics study, focusing on step 4. That is, students in both conditions answered the same essay questions, but students in the non-interaction condition merely read a multi-paragraph explanation, while students in the interaction condition engaged in a typed dialogue with the computer tutor. Contrary to the interaction hypothesis, tutoring did not produce larger gains than reading, although there was a trend in the expected direction.

Lane and VanLehn (in press) compared two versions of a tutoring system that focused on teaching novice programmers to how to design a program before trying to write the code for it. In the interactive conditions, the tutor conducted a typed dialogue with students that elicited a design from them while providing hints and occasional directive help. In the non-interactive condition, students read a text with essential the same content as the tutorial dialogue. Although some post-training measures produced null results, the tutored students exhibited improved ability to compose designs, and their behavior suggested thinking at greater levels of abstraction than students in the reading group. Thus, this experiment supports the interaction hypothesis.

To summarize, there is ample evidence that interaction during tutoring *correlates* with learning gains, which could account for the widespread belief in the learning sciences that interaction *causes* learning gains. However, the correlation could be due to a third factor, such as the students' interest, diligence, etc., that increases both learning gains and interaction. For experiments that compared interactive tutoring to non-interactive instruction, such as lecturing or reading, results varied:

- If students in the comparison condition engaged in no interaction at all and merely read text or sat through a lecture/demonstration, then interactive tutoring elicited larger learning gains than the comparison instruction (Graesser et al., 2001; 2003; Lane & VanLehn, in press; Swanson, 1992; Wood et al., 1978).
- If students in the comparison condition both read text and used the text's content to solve problems or answer questions *during training*, then interactive tutoring was *not* more effective than the comparison instruction (Chi et al., 2001; Jackson et al., 2004; Katz et al., 2003; Rosé et al., 2001; Rosé et al., 2003).

However, null results are often open to many interpretations, and those in the second bullet above are particularly problematic as they are inconsistent with the interaction hypothesis, which is widely believed. In particular, confusing patterns of null and positive results can be caused by aptitude-treatment interactions (ATIs). High-competence students often learn equally well from many types of instructions, whereas low-competence students often learn better from more scaffolded instruction (Cronback & Snow, 1977). When an ATI exists, experiments can have either null results or positive results depending on the prior competence of their students.

In order to test whether the null result, that interactive tutoring tied with mixtures of reading and problem solving, may be due to ATIs or other factors, we conducted 3 experiments. In addition to checking for ATIs, the experiments carefully controlled the mixture of reading and problem solving by varying only step 4 of the 5-step frame. That is, in both the tutoring and comparison conditions, students solved the same training problems (steps 1 and 2 of the 5-step frame). The experimental manipulation affected only the feedback and remediation of the student's solution: the student either interacted with an expert human tutor or read a text. The experiments were conducted as part of a larger study that is reported elsewhere (VanLehn et al, submitted).[1]

Experiment 1: Intermediates

The task domain was qualitative physics. In particular, students were taught to give principle-based answers to essay questions such as:

> "A massive truck has a head-on collision with a lightweight car that is traveling at the same speed. Which vehicle suffers the greater impact force, and which has the greater acceleration? Explain your answers."

An adequate answer to this question would mention both Newton's second and third law:

> "The force exerted by the truck on the car equals the force exerted by the car on the truck, according to Newton's third law. Since the forces are the same, but the truck has a greater mass, the car has a greater acceleration, according to Newton's second law."

Students who have already taken college physics courses often have great difficulty with such questions, in part because they may have misconceptions, such as "heavier objects exert more force" (Hestenes, Wells & Swackhamer, 1992). Cognitive task analyses and simulations (Ploetzner & VanLehn, 1997) suggest that learning of qualitative physics consists of first adapting students' existing equation-based knowledge of principles (e.g., Newton's second law is conceptualized as F=m*a) for use qualitatively, then strengthening these qualitative principles so that they compete successfully with misconceptions. Thus, getting students to apply the principles should suffice to both "compile" them to qualitative forms and to strengthen them.

Thus, our instruction consisted of 10 applications of the 5-step frame, one for each of 10 training questions. In particular, students read a training question, typed in an essay, then engaged in step 4. Students in the Typed Tutoring condition participated in a typed dialogue with an expert human tutor. Students in the Reading condition read a multi-paragraph explanation of the correct answer to the question. When step 4 was completed, students in both conditions read a short, ideal essay similar to the one above, then went on to the next training question.

Student learning was assessed via two tests: (1) one with 4 essay questions similar to the ones used in training, and (2) a 40-question multiple choice test based on the Force Concept Inventory (Hestenes et al., 1992), a standard test of qualitative physics concepts.

Students were recruited from four universities. They were required to have taken college physics, and thus could be considered intermediates. There were 18 students in the Typed Tutoring condition and 22 in the Reading condition.

The main tutor, who tutored about half the students, was a retired physics professor who was employed full-time by the project, had conducted over 170 hours of tutoring during pilot studies of the training materials, and had examined transcripts of his tutoring in order to find ways to improve it. Three other university physics professors also served as tutors. An informal examination of the transcripts showed that lecturing was rare and that the tutoring was highly interactive.

The multiple-choice tests were scored objectively, by counting the number of questions answered correctly. The essay tests were scored in several ways, ranging from a holistic grade (A through F) to a detailed coding of the essays that counted individual correct and incorrect propositions. Table 1 shows the means and standard deviations of the proportion correct on the multiple choice tests and the holistic scoring of the essay tests. ANCOVAs with pre-test scores as covariates showed that although the adjusted post-test scores of the Reading students were slightly higher than the Typed Tutoring students on all measures, none of the differences were statistically reliable.

In order to detect an aptitude-treatment interaction (ATI), students were split into high-pretest and low-pretest groups, and the ANCOVA's were repeated. Neither groups' adjusted post-test scores were reliably different across conditions. An ANOVA that crossed pretest score (high vs. low) with test (pre vs. post) showed no significant interaction.

These finding are consistent with the 5 null results reported earlier (Chi et al., 2001; Jackson et al., 2004; Katz et al., 2003; Rosé et al., 2001; Rosé et al., 2003) and inconsistent with the interaction hypothesis.

Table 1: Experiment 1 means (standard deviations)

	Typed Tutoring	Reading
Multiple-choice pre-test	0.60 (.04)	0.64 (.04)
Multiple-choice post-test	0.74 (.03)	0.79 (.03)
Multi.-choice adjusted post-test	0.77 (.02)	0.79 (.02)
Essay pre-test	0.46 (.05)	0.49 (.05)
Essay post-test	0.68 (.06)	0.74 (.05)
Essay adjusted post-test	0.71 (.05)	0.75 (.04)

Experiment 2: Novices

Although all the students in experiment 1 had taken college physics, some had taken it several years ago. The large but equal gains of the Reading and Typed Tutoring students would be explained if both forms of instruction were equally effective at refreshing their memories. Thus, we repeated the experiment with students who had not taken college physics. (We call these students "novices.") We did not change the training material, which was designed for students who had taken college physics. (We call these students "intermediates.") Thus, we felt it was necessary to pre-train the novices so that they would not be too frustrated by the intermediate-level training. For pretraining, the novices read a short summary of the concepts and principles involved until they felt that they understood it (mean study time: 32 minutes).

We also added a Spoken Tutoring condition while retaining the Typed Tutoring condition. In the Spoken Tutoring condition, students and tutors were seated in the same room and could see the same computer screen with the problem and the student's essay, but a partition prevented them from seeing each other.

[1] Experiments 1, 2 and 3 here correspond to 1, 4 and 5 there.

The main tutor was the only tutor, and students were recruited from only one school (the University of Pittsburgh). Otherwise, the experimental procedure and materials were the same as those in Experiment 1. There were 14 students in the Spoken Tutoring condition, 20 in the Typed Tutoring condition and 20 in the Reading condition.

Once again, students in all 3 conditions had significant learning gains. However, this time an ANCOVA on the multiple-choice post-test scores, using the pre-test as a covariate, showed significant differences between conditions, F(2,50)=10.27, p<.01. The adjusted post-test scores (see Table 2) were ordered Spoken Tutoring > Typed Tutoring > Reading. Pairwise ANCOVAs indicated that the Spoken Tutoring's post-test score was not reliably higher than Typed Tutoring's,, but Typed Tutoring was more effective than Reading (effect size 0.65) and Spoken Tutoring was much more effective than Reading (effect size 1.64).

The essay tests appear not to have been as sensitive as the multiple-choice tests. There were no significant differences among conditions in the adjusted essay post-test scores across all scoring rubrics for essays.

There also was no sign of an ATI using the same median-split analyses as used for ATI in experiment 1. That is, both high-pretest and low-pretest students learned more in the tutoring conditions than in the Reading condition, and by the same amount. Overall, these results support the interaction hypothesis.

Table 2: Experiment 2 means (standard deviations)

	Spoken Tutoring	Typed Tutoring	Reading
MC pre-test	0.42 (.03)	0.46 (.02)	0.44 (.02)
MC post-test	0.74 (.03)	0.67 (.03)	0.57 (.03)
MC adjusted post-test	0.74 (.03)	0.66 (.03)	0.57 (.03)

Experiment 3: Less training

Although using 10 training problems allowed us to detect differences between Reading and Tutoring, such long training required multiple sessions for some students but not others. In experiment 2, students in the Typed Tutoring condition spent on average 180 minutes typing during training and about the same amount of time waiting for the tutor to finish typing. In contrast, students in the Spoken Tutoring and reading conditions completed the training in 165 minutes and 85 minutes, respectively on average.[2] Thus, students in the Typed Tutoring conditions had to return for multiple sessions, which caused higher attrition in that condition than the other conditions.[3] Although there was no difference between the mean pretest scores of the students who dropped out and who stayed in either of the experiments, we decided to repeat the experiment with abbreviated materials and without the Typed Tutoring condition so that all training could be completed in one session.

We removed several principles from the training and the tests. This reduced the training to 5 questions, the multiple-choice test to 26 questions, and the near-transfer essay questions to 3. We added 7 far-transfer essay questions.

We also made several hopefully minor changes: (1) students were recruited from both Pittsburgh and Memphis universities; (2) to accommodate the distance, the Spoken Tutoring condition used telephones; (3) each of the texts of the Reading condition, one per training question, was augmented by adding a summary of the reasoning that was shorter than the full line of reasoning, but longer than the ideal answer presented at the end. This summary was added to make the texts' content more closely approximate the tutorial dialogues' content.

Like experiment 2, the students were novices and studied a short text before the manipulation. There were 21 students in the Spoken Tutoring condition and 19 in the Reading condition. All students completed the experiment in one session, and none dropped out of the experiment.

Once again, we found students in both conditions had large gains between pre- and post-tests (see Table 3, upper half). However, for all tests and all scoring rubrics, there were no reliable differences between the adjusted post-test scores of the Spoken Tutoring students and the Reading students, although there were trends in the expected direction. Thus, the results do not support the interaction hypothesis.

However, we did find an ATI (see Table 3, lower half). Among the low-pretest students, the Spoken Tutoring students had higher adjusted post-test scores on the multiple-choice test than the Reading students, F(1,17)=5.876, p<.03. Among the high-pretest students, this difference was not statistically reliable.

Table 3: Experiment 3 means (standard deviations)

	Spoken Tutoring	Reading
Multiple-choice pre-test	0.49 (.04)	0.41 (.04)
Multiple-choice post-test	0.68 (.04)	0.56 (.04)
Multi.-choice adjusted post-test	0.66 (.03)	0.60 (.03)
Low pre-test: MC pre-test	0.32 (.06)	0.30 (.06)
Low pre-test: MC post-text	0.65 (.10)	0.46 (.19)
Low pre-test: MC adj. post-test	0.65 (.16)	0.46 (.16)

Discussion

First, let us review the findings. In all 3 experiments, students were asked training questions, provided a paragraph-long typed explanation, and then either read a text explaining the correct answer (the Reading condition), or interacted with an experienced human tutor either orally (the Spoken Tutoring condition) or via typing (the Typed Tutoring condition). In Experiment 1, which used students who had already taken college physics, Reading students tied with Typed Tutoring students. In experiment 2, which

[2] That Spoken Tutoring is more effective than Reading in experiment 2 is probably not due solely to time-on-task, because Typed Tutoring took even longer than Spoken Tutoring and yet had lower gains.

[3] In experiment 1, 3 of 21 students dropped out of the Typed Tutoring condition, and 3 of 26 dropped out of the Reading condition. In experiment 2, 5 of 25 dropped out of Typed Tutoring, 3 of 17 dropped out of Spoken Tutoring, and 0 of 20 dropped out of Reading.

used students who had not taken college physics but had read a short summary of the relevant concepts and principles, Typed Tutoring and Spoken Tutoring students learned more than Reading students. In experiment 3, which used students who had not taken college physics and used half as many training problems, Spoken Tutoring students tied with Reading students. However, for low-pretest students in experiment 3, Spoken Tutoring students learned more than Reading students. No similar ATIs were observed in experiments 1 and 2.

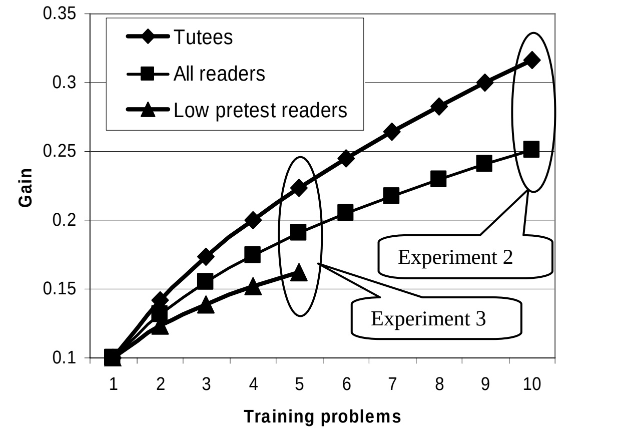
When these results are put in the context of earlier research, a clear pattern emerges:

- Although texts cannot adapt to the competence of the reader, human tutors can adapt to their tutees. Thus, when a text is too difficult for a student, tutoring will be more effective than reading. This explains why Spoken Tutoring was more effective than Reading in experiment 2.
- If the training material is at the right level for the students, but the students are not required to use its content during training, then they may not be fully engaged in the reading or may not set their metacognitive standards of comprehension at a sufficiently deep level. This explains why interactive tutoring was more effective than reading that is not accompanied by problem solving or question answering during training (Graesser et al., 2001; 2003; Lane & VanLehn, in press; Swanson, 1992; Wood et al., 1978).
- If the training material is at the right level for the students and the students are required to use its content during training, then interactive tutoring is no more effective than mixing reading with problem solving. This explains our experiment 1 findings and 5 others (Chi et al., 2001; Jackson et al., 2004; Katz et al., 2003; Rosé et al., 2001; Rosé et al., 2003).

The results of our experiment 3 are consistent with this pattern as well, albeit somewhat more complicated to explain. Suppose that the learning rate in the Reading condition was less than the learning rate in the Spoken Tutoring condition, because the text was over the students' heads (see top two curves of Figure 1). Because the Spoken tutees learned faster than the Readers, the longer the training, the further apart the two groups will become. In experiment 2, their learning gains at the end of the 10 training problems were far enough apart to be statistically reliable, but in experiment 3, the learning gains at the end of 5 training problems were not far enough apart to be reliably different. This explains the main effect of experiment 3, which is that the Spoken Tutoring condition was not more effective than the Reading condition.

In order to explain the ATI of experiment 3, we can assume that the low pre-test students had even lower learning rates from reading than the high pre-test students. That is, the text far over the heads of the low-pretest students, so the lowest curve on Figure 1 represents their learning. Let us also assume that the human tutor was able to adapt equally well to the low and high pre-test students, so their learning rates were equal (top curve). Thus, even though experiment 3 had only 5 training problems, the extra low learning rate of the low pre-test readers insures that their gains will be far enough apart from the gains of the low pre-test tutees to be statistically significant.

Figure 1: Hypothetical learning curves of novices



To put it more concretely, suppose the student is asked, "A massive truck has a head-on collision with a lightweight car that is traveling at the same speed. Which vehicle suffers the greater impact force?" The student answers, "The truck exerts a greater impact force on the car because it has a larger mass." First consider how the tutor might discuss this answer with the student:

- Tutor: So you think the truck's force is larger?
- Student: Yes
- Tutor: Well, consider that the truck is exerting a force on the car, but the car is also exerting a force on the truck. You've got two forces involving the same objects. Does this remind you of any laws?
- Student: Newtons' first law?
- Tutor: Try third. Do you know what it is?
- Student: For every force of A on B, there is an equal and opposite force of B on A.
- Tutor: Excellent! Can you apply that here?
- Student: The force of the truck on the car is equal and opposite the force of the car on the truck.
- Tutor: Great!

Now consider what is learned if the student reads the following paragraph instead of participating in a tutorial dialogue:

> "As we know from Newton's third law, when two objects exert forces on each other, the forces have equal magnitudes. Thus, the force exerted by the truck on the car must equal the force exerted by the car on the truck."

If the student pays attention when reading the paragraph and has no comprehension difficulties due to lack of prior knowledge, then it seems plausible that the student would learn just as much from reading the paragraph as from participating in the tutorial dialogue. Although the dialogues and texts in our experiments were much longer than the ones above, it is still intuitively plausible that the learning gains would be the same provided that students had appropriate background knowledge and that they were motivated to pay close attention to the text because they knew that they would soon have to answer more questions.

These remarks are quite speculative and more experimental work is clearly needed. Nonetheless, it is safe to say that the common belief that interactive instruction is always better than non-interactive instruction is probably unwarranted given our results and those from earlier work. Even two extreme forms—reading vs. one-on-one interactive spoken human tutoring—can produce the same gains under certain conditions, such as our experiment 1. These results are important not only for developers of computer tutoring systems, but for the whole debate over constructivist vs. didactic instruction.

Acknowledgments

This research was supported by grants from DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research (N00014-00-1-0600) and from the National Science Foundation (SBR 9720314, REC 0106965, ITR 0325428, NSF 9720359 and EIA-0325054). We would like to thank faculty, staff, and students in the Tutoring Research Group (TRG) at the University of Memphis and the Natural Language Tutoring group at the University of Pittsburgh.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

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

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

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

Graesser, A.C., Person, N., Harter, D., & TRG (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 257-279.

Graesser, A.C., Jackson, G.T., Mathews, E.C., Mitchell, H.H., Olney, A., Ventura, M., Chipman, P., Franceschetti, D., Hu, X., Louwerse, M.M., Person, N.K., & TRG (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman and D. Hirsh (Eds.), *Proceedings of the 25rd Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Katz, S., Connelly, J., & Allbritton, D. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence in Education, 13*, 79-116.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30*, 141-158.

Jackson, G.T., Ventura, M.J., Chewle, P., Graesser, A.C., and the Tutoring Research Group (2004). The impact of Why/AutoTutor on learning and retention of conceptual physics. In J.C. Lester, R.M. Vicari, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems 2004* (pp. 501-510). Berlin, Germany: Springer.

Lane, H. C., & VanLehn, K. (in press). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education*..

Ploetzner, R., & VanLehn, K. (1997). The acquisition of informal physics knowledge during formal physics training. *Cognition and Instruction, 15*(2), 169-206.

Rose, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2001). A comparative evaluation of Socratic versus didactic tutoring. In J. D. Moore & K. Stenning (Eds.), Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society (pp. 897-902). Mahwah, NJ: Erlbaum.

Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & Vanlehn, K. (2003). Exploring the effectiveness of knowledge construction dialogues. Paper presented at the AIED, Sydney, Australia.

Swanson, J. D. (1992) What does it take to adapt instruction to the individual?: A case study of one-to-one tutoring. Paper presented at AERA, San Franscisco, CA.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A. & Rosé, C.P. (submitted) Natural Language Tutoring: A comparison of human tutors, computer tutors and text.

Wood, D., Wood, H. & Middleton, D. (1978). An experimental evaluation of four face-to-face teaching strategies. *International Journal of Behavioral Development, 1*, 131-147.