Is Human Tutoring Always More Effective than Reading?: Implications for Tutorial Dialogue Systems

Carolyn P. Rosé, Kurt VanLehn, and the Natural Language Tutoring Group
Learning Research and Development Center, University of Pittsburgh, Pittsburgh PA, 15260

Abstract. Some previous studies of student learning have demonstrated a strong advantage in favor of human tutoring over a classroom control condition (Bloom, 1984; Cohen et al., 1982). These results have spawned an optimistic view towards building effective tutorial dialogue systems. Towards this end, many current tutorial dialogue systems have been evaluated successfully with students [21, 15, 1, 11, 8]. Nevertheless, so far none have demonstrated conclusively that tutorial dialogue systems provide a more effective or efficient means of instruction than an otherwise equivalent purely text based approach. In this paper we explore the question of whether it is even true that human tutoring is always superior to a reading control.

In recent years much attention has been given to questions about tutorial dialogue, in particular about what makes it effective and in which contexts. The current study was motivated by the hypothesis that dialogue may be a more appropriate means of instruction for naive learners than for review learners. In this study review learners are those who have been exposed to the material in a formal classroom setting but have not yet mastered the material. Our study focuses on learning conceptual physics. We used two different populations of students. In particular, the first population of students were those who were in the middle of taking or had already taken a semester of college level physics. Thus, for these students the topics covered in the study were a review of what they had already learned in class but did not yet master (as indicated by pretest scores). The second population of students were those who had never taken any college level physics. While students in both conditions demonstrated a significant difference in performance between the pre-test and post-test, there was no significant difference between conditions with the first population, whereas there was a significant difference in gain between the human tutoring condition and the reading control with the second population. This interaction supports the experimental hypothesis and highlights the benefit of adaptation to student knowledge state that dialogue affords. These results also have methodological implications for tutorial dialogue research.

1 Introduction

Previous studies of student learning have provided compelling evidence that one-on-one human tutoring is more effective than other modes of instruction (Bloom, 1984; Cohen et al., 1982). Since human tutoring is normally conducted through natural language dialogue, perhaps in conjunction with other communication modalities, it has been conjectured that it is to a large extent, the collaborative dialogue between student and tutor that makes human tutoring such a powerful vehicle for instruction (Fox, 1993; Graesser et al., 1995; Merrill et al., 1992, Chi et al., 2001).

Tutorial dialogue has many positive characteristics for promoting learning. First, it is a natural way to provide students with a learning environment that exhibits characteristics that have been shown to correlate with student learning gains, such as student activity. For example, it has been demonstrated that generating words rather than simply reading them promotes subsequent recall of those words (Slamecka and Graf, 1978). Furthermore, encouraging student self-explanation, which includes both generating inferences from material they have read and relating new material to old material, has been shown to correlate with learning (Chi et al., 1984; Renkl, 1997; Pressley et al., 1992). In a further study, prompting students with zero content prompts to encourage them to self-explain was also associated with student learning (Chi et al., 2001).

A second important advantage to dialogue is that it affords the tutor the opportunity to tailor instruction to the needs of the student. While human tutors may not always choose to tailor their instruction to the individual characteristics of the knowledge state of their students, in this study we highlight the benefits of adaptation by comparing human tutoring to a completely non-adaptive reading control. If adaptation in dialogue is a significant factor in its relative effectiveness with other modes of instruction, then one would expect that reading might be just as effective for instruction if the content of the text that was presented to the student could be tailored appropriately to the student's knowledge state. Such tailoring is only possible with a relatively uniform population of students, for example, one with a large shared base of relevant knowledge and experience. Thus, we contrast two populations, one of which is relatively uniform in this sense and one of which does not. In particular, we contrast review learners who are in the middle of or have already completed their first college level physics course with students who have never had any formal college level physics instruction. Because review learners have already received formal instruction on the relevant topics, they share a large base of relevant information that they have all been exposed to in a rigorous manner. Naive learners, on the other hand, have spotty knowledge about the relevant topics.

The observed treatment interaction makes sense when one considers both the pros and cons of tutorial dialogue. While tutorial dialogue has many benefits for the student, it comes with a cost. In particular, engaging in a dialogue with a human tutor about a topic requires more time and energy from the student than reading a summary of the same material. In order for tutorial dialogue to be an appropriate means of instruction, the benefits must outweigh the costs. Because adaptation in tutorial dialogue may not be as beneficial to a population of review learners, they may not benefit as much from tutorial dialogue in comparison with a population of naive learners. Furthermore, the benefit of dialogue in comparison with a reading control, where the material is targeted appropriately to their knowledge state, may not outweigh the cost.

Here we test the hypothesis that dialogue may be a more appropriate means of instruction for naive learners than for review learners. We ran the study in two parts. Specifically, we ran the review learners in the Spring of 2002 and the naive students in the Fall of 2002. We describe these studies and explore the implications that these results have on tutorial dialogue systems research.

2 The Spring 2002 Study

In the Spring of 2002, we ran a study to compare the effects of human tutoring with a reading control, using qualitative physics as the task domain. In both conditions, the activity in

Student1 Both balls will hit the ground at the same time. The balls are in free fall (only gravitational forces). The ratio of the masses and weight are equal.

Tutor1: You have correctly said the balls are in free fall. What do you conclude from this fact?

Student2: There is no air resistance and the balls' accelerations are constant in regards to one another

Tutor2: Right, so the conclusion is that they both have the same acceleration, the free fall acceleration. Now you have to show that time taken by them to reach the ground will be the same. How do you do that?

Student3: F (net forces) / mass = f/m because F = m*a therefore ratio should equal one another

Tutor3: But that does not tell you anything about the time taken. What physical quantities will determine the time taken. Will the initial velocity of the balls matter?

Student4: yes, assuming starting at rest? sorry, kind of at a loss at the moment

Tutor4: The initial velocity will matter, and here they both have the same initial velocity, zero. What else, will the height through wich they fall matter?

Student5: not so much as long as they are both dropped from the same height

Tutor5: Height of fall will determine the time of fall, and here the height of fall is the same for both the balls. Now on the basis of these facts can you conclude about the time taken by them to reach the ground?

Student6: since initial velocities are equal and the height of falls are equal, then the time taken is equal

Tutor6: How about acceleration, if they had different accelerations, even if they fall through the same height and have same initial velocity , will they reach the ground at the same time?

Student7: no...

Tutor7: Right...

Figure 1: **Sample tutoring dialogue**

which students engaged was answering deep reasoning questions involving topics in conceptual physics. One such question that we used is, "A lightweight car and a massive truck have a head-on collision. On which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Explain." This is an appropriate task domain for pursuing questions about the benefits of tutorial dialogue for learning because questions like this one are known to elicit robust, persistent misconceptions from students, such as "heavier objects exert more force." (Hake, 1998; Halloun and Hestenes, 1985), which is a known common misconception about Newton's Third Law. We designed a set of 10 essay questions to use for instruction. Two physics professors and a computer science professor worked together to select a set of expectations (i.e., correct propositions that the tutors expected students to include in their essays) and potential misconceptions associated with each question. Additionally, they agreed on an ideal essay answer for each problem.

The tutors were instructed to cover the expectations for each problem, to watch for a specific set of expectations and misconceptions associated with the problem, and to end the

You may have heard about Galileo's experiment of dropping two fairly heavy balls of different masses from the top of the leaning tower of Pisa in Italy. He observed that each ball took to the same time to strike the ground. In terms of the principles of physics the forces acting on these balls are earth's gravity and air resistance.

To analyze the motion of any object one must recognize the forces acting on it. Earth's gravity and air resistance are acting on the balls. The force of air resistance is always present if motion is through the air, but it is very small as compared to the force of earth's gravity on the balls if their speed is not large. This force, in most cases, can be neglected. If air resistance is neglected, then the only the force acting on each ball is the force of earth's gravity. When an object moves with the earth's gravity being the only force acting on it, its motion is called freefall.

Now consider Galileo's observation. The balls of different masses, dropped (i.e.they had the same, zero initial velocity), from the same height take the same time to reach the ground. From this we can conclude that their velocities at any time during the fall are the same at every time. So, their acceleration during this freefall is the same. Of greater significance is that acceleration during freefall is the same for all objects and it does not depend on an object's mass. This freefall acceleration is due to gravity.

The value of 'g' varies very little on the surface of the earth as well as at ordinary heights that we come across while considering the motion of objects near earth. Thus, can we take it as a constant and do NOT need to consider it as varying at or near the earth. The commonly accepted approximate value of 'g' on earth, i.e., acceleration due to earth's gravity or freefall acceleration is 9.8m/s/s.

Figure 2: **Sample minilesson**

discussion of each problem by showing the ideal essay to the student. They were encouraged to avoid lecturing the student and to attempt to draw out the student's own reasoning. They knew that transcripts of their tutoring would be analyzed. Nevertheless, they were not required to follow any prescribed tutoring strategies. So their tutoring style was much more naturalistic than in previous studies such as the BEE study (Rosé et al., 2000) in which two specific tutoring styles, namely Socratic and Didactic, were contrasted. The results of that study revealed a trend for students in the Socratic condition to learn more than those in the Didactic condition. A further analysis of the corpus collected during the BEE study (Core et al., 2002) verified that the Socratic dialogues from the BEE study were more interactive than the Didactic ones. The biggest reliable difference between the two sets of tutoring dialogues was the percentage of words spoken by the student. The Didactic dialogues contained on average 26% student words, whereas the Socratic dialogues contained 33% student words. On average with respect to percentage of student words, the dialogues in our corpus were more like the Socratic dialogues from the BEE study, with average percentage of student text being 31%. Nevertheless, because the tutor was not constrained to follow a prescribed tutoring style, the level of interactivity varied widely throughout the transcripts, at times being highly Socratic, and at other times being highly Didactic. A sample tutoring dialogue from our corpus is displayed in Figure 1. Here we see our tutor employing a highly interactive tutoring style, where tutor turns are relatively short and contain a large number of probing questions.

In both conditions, the students were presented with as many of 10 essay questions as they were able to work through in 8 hours, which was divided into sessions of not more than 4 hours each. For each question, the student first wrote an initial essay, received some instruction, and then revised the essay. In the human tutoring condition the instruction was in the form of a dialogue between the student and the tutor through a text-based chat interface with student and tutor in separate rooms. At key points in the dialogue, the tutor asked the student to revise the essay. This cycle of instruction and revision continued until the tutor was satisfied with the student's essay. In the reading control, after the student completed an initial essay attempt, the student was presented with a set of minilessons. For each problem there was one minnilesson targeting each of the required points for the problem, and one targeting a correct understanding of a conceptual physics topic corresponding to each of the key misconceptions associated with the problem. A sample minilesson is displayed in Figure 2. After reading the full set of minilessons, which covered all of the conceptual knowledge and lines of reasoning required to write a completely correct essay answer, the student then revised the initial essay. When students completed the instruction and revised their essay a final time, they then read the ideal essay associated with the problem.

The minilessons were intended to contain all of the inferencing required to correctly solve the problem. Unlike a text book, where a much greater volume of information is presented, only some of which is relevant to the task at hand, the minilessons provide the student with exactly what is required to perform the task of constructing a complete and correct essay with perfect understanding. Thus, it is a control condition that is very hard to beat. Minilessons are written in a simple style with reference to scenarios that should be easy for students with some physics background to understand. Key concepts such as Newton's laws were repeated often. Nevertheless, a student encountering the information for the first time may have trouble absorbing it without any coaching.

2.1 Experimental Design

The complete study run in Spring 2002 was a 4-way design with the two conditions we discuss here as well as two additional conditions in which students interacted with two different prototype systems designed to mimic the behavior of the human tutoring condition (Graesser, VanLehn, TRG, and NLT Group, 2002). Students were randomly assigned to one of the four conditions. 96 students completed the study, 40 of which were in the two conditions we are concerned with in this paper. 20 students started the Human tutoring condition, 18 of which finished. 24 students started the minilesson condition, 22 of which finished. In neither case was there a significant difference between the average pretest score of the students who remained in the study and those who did not.

The subjects were university students who were currently taking or had recently taken introductory college physics and had not taken advanced physics courses or mechanical engineering courses. If the students were currently taking college physics, then they were required to have taken their first midterm because it covered main topics of the tutoring (kinematics and forces). The subjects were volunteers responding to advertisements at the University of Pittsburgh, the University of Memphis and Rhodes College. Students were compensated with money or extra course credit. Four physics professors participated in the study as tutors, although one of the four professors tutored half of the subjects in the human tutoring condition, where the other students were split between the remaining three tutors. Each student was tutored by only one of the four tutors.

Two tests were developed as pre/post tests: versions A and B, which were isomorphic to one another. That is, the problems on test A and B differed only in the identities of the objects (e.g., cars vs. trucks) and other surface features that should not affect the reasoning required to solve them. Each version of the test (A and B) consisted of 40 multiple choice questions. Each multiple choice question was written to address a single expectation covered in the training problems. Some students were not able to complete all 10 problems before they reached the end of their participation time. Thus, they took the post-test after only working through a subset of the training problems.

2.2 Results

Test items were scored as right or wrong; no partial credit was given. In our statistical analyses, we adopted an alpha of .05 in tests of statistical significance. Our analyses showed no significant difference between conditions.

A t-tailed unpaired t-test demonstrates that the pretest scores were not reliably different between the two conditions. Mean pretest score in the human tutoring condition was .60 with a standard deviation of .17, whereas the mean pretest score in the minilesson condition was .64 with a standard deviation of .18. t(38)=0.77; p=0.44. So the students in both conditions started out on an even playing field with respect to incoming competence. There was also no significant difference between average posttest scores. Mean posttest score in the human tutoring condition was .79 with standard deviation of .14. Mean posttest score in the minilesson condition was .74 with a standard deviation of .14. A 2-tailed unpaired t-test showed no significant difference. t(38) = 1.08; p=.29. Nevertheless, students in both conditions learned a significant amount during their involvement in the study, on average gaining about 15% between pretest and posttest.

In summary, the students in both groups learned, and learned about the same amount, at least on concepts covered by the pretest and posttest. While the human tutors were instructed to emphasize the same concepts covered in the minnilessons, they did sometimes discuss topics that were not covered, such as definitions of basic terms. Thus, students in the human tutoring condition may have learned additional information that students in the minnilesson condition did not have the opportunity to learn. However, those topics were raised in support of the topics covered on the test. And what we are interested in for this study is supporting learning of the topics that are covered on the test. What we conclude from our analysis is that whatever benefit students received from the additional information and the interactive nature of dialogue was not enough to lead to significantly increased learning of the topics we were concerned with.

3 Fall 2002 Study

In the Fall of 2002, we reran the same study, but this time with students who had never taken any college level physics courses. The subjects were all University of Pittsburgh students. This time only one tutor participated in the study, in particular the same tutor who tutored half of the human tutoring subjects from the Spring study. The experiment took altogether no more than 15 hours per student, which was split up between 1 and 5 sessions at the end of the Fall 2002 semester, continuing in the Spring 2003 semester. 20 students completed both the minilesson condition and the human tutoring condition.

All of the same materials used in the Spring 2002 study, including the problem set, the minilessons, and the tests, were reused in the Fall 2002 study. Because in this study we used subjects who had never taken college level physics before, after taking the pretest, subjects then read through a 9-page document summarizing some physics background material that would be helpful to them as they worked through the 10 training problems. This material was extracted from chapters 2-6 of (Hewit, 1987).

3.1 Results

Tutor1: You have used the 3rd law correctly. But how did you infer that the car will have greater change in motion?

Student1: First of all, I took a guess that they would be going about the same speed, so that wouldn't really matter, alot like my last senetence of the essay, so I figured that the truck would have a whole lot more mass behind it, and thus more inertia and such, so it would beable to move the little car quite a distance.

Tutor2: Speed is not needed, that is why it is not mentioned in the question. One vehicle could be parked the other moving or both could be moving with varying speeds. Now that you have used the 3rd law and rightly inferred that the forceof impact on either vehicle will have the same magnitude, can you use the 2nd law to determine which will have greater acceleration?

Student2: Sure, the force is the same, the truck is heavier, so it would show that the car is going to have the greater acceleration.

Tutor3: Two objections. Forces is NOT the same, they are of the same magnitude. Same forces will require same magnitude and same directions, as force is a vector. Second, heavier involves comarioson of weights but the 2nd law requires mass not weight. So you can say more massive but heavier will be inaccurate.

Figure 3: **Example of Tutor Adaptation**

A comparison between pre and post tests between conditions in this study shows a significant difference in favor of human tutoring. 45 students began the study, with 40 finishing. 25 students began the human tutoring condition with 20 finishing. There was no significant difference between average pretest scores of those who finished and those who did not. All 20 students who began the reading condition finished the study. There was no significant difference between average pretest scores of students in the two conditions. In particular, mean pretest score for human tutoring students was 0.46 with a standard deviation of .09, whereas mean pretest score for the minilesson condition was 0.44 with a standard deviation of .11.

However, there was a significant difference between posttest scores. Mean posttest scores for the minilesson condition was 0.57 with standard deviation .13, whereas mean posttest score for the human tutoring condition was 0.67 with a standard deviation of .13. t(38)=2.6; p=.01; effect size = 0.70 sigmas.

We conducted a further investigation to begin to uncover which aspects of the tutorial dialogue were responsible for its effectiveness, and in particular to measure the extent to which the tutor adapted the presentation of material to the needs of the student. The first piece of evidence that the tutor was attending to and responding to the contributions of the student was that the average length of tutor turns within the transcripts for each student were highly correlated with the average length of the student turns. A linear regression with average student turn length as the independent variable and average tutor turn length as the dependent variable showed a reliable correlation (R=.565;p<.05).

This correlation by itself is not enough to demonstrate that the tutor is adapting to the student, since it could also mean that students say more when tutors say more. However, in an analysis of the BEE dialogues (Core et al., 2002) it was demonstrated that when tutors adopted a more Didactic, lecture like style, students said less. An alternative explanation is that tutors who *attend to* and *respond* to the contributions of their students say more when students say more because there is more to respond to. Informally, we have frequently observed in our corpus patterns similar to that displayed in Figure 3. Here we see in `Tutor1:` the tutor beginning with a probing question about how a student came to a particular conclusion found in his essay. The student responds in `Student1:` with a lengthy explanation. In `Tutor2:` the tutor points out both some correct and incorrect aspects of the student's explanation. Then the tutor builds on this with a further question. The student then responds, and the tutor finally offers some final corrections. We plan to conduct a more formal analysis of our corpus to determine precisely the distribution of interactions like this one and how they correlate with student learning. Some preliminary evidence suggests that students offer more lengthy responses to tutor turns when they receive more explicit *feedback* on what they have said [20].

Longer student answers to tutor questions reveal more of a student's reasoning. Very short answers, i.e., 10 words or less, are normally composed of a single clause at most. Longer, multi-clausal answers have the potential to communicate many more inter-connections between ideas. Thus, if a tutor is attending to and responding directly to the student's revealed knowledge state, it would be expected that the effectiveness of the tutor's instruction would increase as average student turn length increases. To test this prediction, we computed a linear regression of the sequence of student turn lengths over time for each student in order to obtain an intercept and a slope. We use the intercept in place of the mean length since we have observed a trend for student answers to decrease in length over the course of their interaction with the tutor. We then computed a multiple regression with pre-test score, intercept, and gradient as independent variables and post test score as the dependent variable. We regressed out pretest score in addition to slope since students with higher pretest scores naturally achieve higher post-test scores. We found a reliable correlation between intercept and learning, with pre-test scores and gradients regressed out, both within the naive learner population (R=.836; p<.05) and within the review learner population (R=.454; p<.05), although the correlation is stronger within the naive learner population where the effect of human tutoring was more noticeable. This result is consistent with (Core et al., 2002).

Some methodological issues related to the current studies must be acknowledged. The student population used in the Fall2002 study was not primarily from the same university as that used in the Spring2002 study, which could have influenced the outcome of the experiment. Secondly, the second study used only a single human tutor, whereas the first study used four, although half of the students from the first study were tutored by the same tutor used in the second study. We cannot rule out the possibility that the difference in outcome between the two studies described here could have been a result of these two factors.

4 Current Directions

The results of the current study are consistent with the hypothesis that human tutoring is as effective as it is because the tutor can adapt the presentation of material more directly to the student's needs. Thus, if review learners have sufficient background to grasp the material as it is presented, then the material is already tailored to their knowledge state. For these students, engaging in a dialogue may not be more beneficial than reading the text. Conversely, if tutors do not tailor their discussion of material to the student's knowledge state as it is revealed in the dialogue, then engaging in that dialogue may be no more effective for instruction than reading a text that covers the same material. This is in line with previous claims that tutors who ignore signs of student confusion may run the risk of preventing learning (Chi, 1996).

One might argue that an alternative possible explanation for the difference between the review learner population and the naive learner population is that the instruction simply reminded the review learners of what they already knew and that they did not learn any new knowledge. Since their participation in the study simply returned them to their previous level of physics understanding, any form of instruction would have sufficed. This could be put forward as an explanation of the null effect in the first study as an argument against the hypothesis that tutor adaptation is the important factor. However, this explanation would suggest that the instruction from the human tutors had nothing to offer review learners despite the fact that they demonstrated a lack in areas specifically addressed in the training problems. Thus, it would leave the question open of why review learners would not learn from their interaction with the tutors when they received feedback from them about the parts of the relevant knowledge that their explanations revealed to be lacking. Furthermore, it does not offer an explanation about the difference between conditions with naive learners.

Recent studies of human tutoring suggest that a productive activity for teaching is to have students explain physical systems qualitatively (Chi et al., 1984). An additional possible explanation for the difference between conditions with naive learners is that students in the read only condition did not self-explain except in their initial essays. Although students in the read only condition revised their essays after reading the minilessons, their final essays could be considered simply to be paraphrases of the minilessons they read since all of the argumentation required to solve the problem were found in the minilessons presented to the students. This alternative explanation also seems unsatisfying since it seems inconsistent with the fact that students in the reading condition improved their performance in between the pretest and posttest in both studies.

The results from this study have implications for tutorial dialogue systems research. In particular, while the results from the often cited Bloom and Cohen studies (Bloom, 1984; Cohen et al., 1982) have lead to a strong optimism about the potential of tutorial dialogue systems, we are yet to see convincing proof that tutorial dialogue systems provide a more effective or efficient means of instruction than an otherwise equivalent purely text based approach. The results of the studies reported here cast doubt that even human tutoring is always superior to a well crafted reading control. While much prior research has explored common patterns of human tutorial dialogue, very few of these studies have explored correlations between these patterns found and student learning. We argue that before we will be able to build tutorial dialogue systems that are provably superior to informationally equivalent reading controls, we must determine where ideal tutorial dialogue leads to the greatest advantage and what characteristics of this dialogue are most responsible for this advantage.

5 Acknowledgments

This research was supported by the Office of Naval Research, Cognitive Science Division under grant number N00014-0-1-0600 and by NSF grant number 9720359 to CIRCLE, Center for Interdisciplinary Research on Constructive Learning Environments at the University of Pittsburgh and Carnegie Mellon University.

References

[1] K. D. Ashley, R. Desai, and J. M. Levine. Teaching case-based argumentation concepts using didactic arguments vs. didactic explanations. In *Proceedings of the Intelligent Tutoring Systems Conference*, pages 585–595, 2002.

[2] B. S. Bloom. The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984.

[3] M. Chi, N. de Leeuw, M. Chiu, and C. LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 1984.

[4] M. T. H. Chi. Learning processes in tutoring. *Applied Cognitive Psychology*, 10:S33–S49, 1996.

[5] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann. Learning from human tutoring. *Cognitive Science*, (25):471–533, 2001.

[6] Peter A. Cohen, James A. Kulik, and Chen-Lin C. Kulik. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19:237–248, 1982.

[7] M. Core, J. D. Moore, and C. Zinn. Initiative in tutorial dialogue. In *Proceedings of the ITS Workshop on Empirical Methods for Tutorial Dialogue Systems*, pages 46–55, 2002.

[8] M. Evans and J. Michael. *One-on-One Tutoring by Humans and Machines*. Lawrence Earlbaum and Associates, in-press.

[9] Barbara A. Fox. *The Human Tutorial Dialogue Project: Issues in the design of instructional systems*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.

[10] A. Graesser, K. Vanlehn, TRG, and NLT Group. Why2 report: Evaluation of why/atlas, why/autotutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Technical report, LRDC Tech Report, University of Pittsburgh, 2002.

[11] A. Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz, and the Tutoring Research Group. Autotutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1(1):35–51, 1999.

[12] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. Collaborative dialogue patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology*, 9:495–522, 1995.

[13] R. R. Hake. Interactive-engagement versus traditional methods: A six-thousand student survey of mechanics test data for introductory physics students. *American Journal of Physics*, 66(64), 1998.

[14] I. A. Halloun and D. Hestenes. The initial knowledge state of college physics students. *American Journal of Physics*, 53(11):1043–1055, 1985.

[15] N. T. Heffernan and K. R. Koedinger. An intelligent tutoring system incorporating a model of an experienced tutor. In *Proceedings of the Intelligent Tutoring Systems Conference*, pages 596–608, 2002.

[16] P. G. Hewitt. *Conceptual Physics*. Adison Wesley, 1987.

[17] Douglas C. Merrill, Brian J. Reiser, and S. Landes. Human tutoring: Pedagogical strategies and learning outcomes, 1992. Paper presented at the annual meeting of the American Educational Research Association.

[18] M. Pressley, E. Wood, V. E. Woloshyn, V. Martin, A. King, and D. Menke. Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27:91–109, 1992.

[19] A. Renkl. Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21(1):1–29, 1997.

[20] C. P. Rosé, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn.

[21] C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. Interactive conceptual tutoring in atlas-andes. In *Proceedings of Artificial Intelligence in Education*, pages 256–266, 2001.

[22] C. P. Rosé, J. D. Moore, K. VanLehn, and D. Allbritton. A comparative evaluation of socratic versus didactic tutoring. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 869–874, 2001.

[23] N. J. Slamecka and P. Graf. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, (4):592–604, 1978.