Adding Abstractive Reflection to a Tutorial Dialog System

Arthur Ward
University of Pittsburgh
Department of Biomedical Informatics, Suite 301
5150 Centre Avenue
Pittsburgh, Pa., 15232, USA
akw13@pitt.edu

Diane Litman
University of Pittsburgh
Learning Research and Development Center
Department of Computer Science
Pittsburgh, Pa., 15260, USA
litman@cs.pitt.edu

Abstract

In this work we hypothesize that giving students a reflective reading after spoken dialog tutoring in qualitative physics will improve learning. The reading is designed to help students compare similar aspects of previously tutored problems, and to abstract their commonalities. We also hypothesize that student motivation will affect how well the text is processed, and so influence learning.

We find that the beneficial effects of the reflective text significantly interact with motivation, such that moderately motivated students learn significantly more from the reflective text than from a non-reflective control text. More poorly or highly motivated students did not benefit from reflective text.

These results demonstrate that implicit reflection can improve learning after dialog tutoring with a qualitative physics tutor. They further demonstrate that this result can be obtained with a reflective/abstractive text without recourse to dialog, and that the effectiveness of the text is sensitive to the motivation level of the student.

Introduction

Researchers have been aware for some time that students can learn to solve numerical physics problems while still retaining a poor overall knowledge of physics concepts (Halloun and Hestenes, 1985b,a). A number of tutoring system projects, such as Atlas-Andes (Rosé et al., 2001), Why2-Atlas (VanLehn et al., 2002), AutoTutor (Graesser et al., 2005), and Itspoke (Litman and Silliman, 2004), have attempted to address the problem of poor conceptual learning by adding natural language instruction to quantitative physics problem solving tutors, or by using natural language dialog to teach physics concepts directly. We will use the term “quantitative” to refer to a tutoring system which emphasizes solving numerical physics problems. Systems which attempt to teach physics concepts without numerical problem solving we will call “qualitative.”

All of these systems have produced learning gains. In general, however, the post-test scores are still fairly low. For example, average correctness on a post-test administered during a 2008 study with the Itspoke tutor was only about 74%. Therefore, in this work we seek to further improve students’ conceptual learning from using a qualitative physics tutor. Our approach to this is motivated by the literature in transfer and reflection.

In their work on transfer, Gick and Holyoak (1983) investigated how learning a source problem could improve performance on an analogous target problem. They found that many interventions, such as having the student memorize, summarize or diagram the source problem, had no effect on transfer. However, when the students were asked to compare two source problems, performance on the target problem was (finally) improved. They hypothesized that comparing these source analogs resulted in the generation of more abstract and transferable problem representations.

As will be described in more detail below, the Itspoke tutor engages students in dialogs about several different problems in conceptual physics. In these problems, the same physics concepts are applied in different physical situations. Gick and Holyoak’s result raises the question of whether a process of comparison and abstraction between these concept applications would help students better learn physics concepts from the Itspoke tutor.

This comparison of different tutoring dialogs would necessarily be done reflectively, after they had been completed. Several types of reflection have been studied in the literature, with a distinction sometimes drawn between “reflection-in-action,” which happens during problem solving, and “reflection-on-action” which happens after problem solving (Tchetagni, Nkambou, and Bourdeau, 2007). We examine the reflection-on-action literature to see how it could encourage comparison and abstraction.

In several important studies (Katz, Allbritton, and Connelly, 2003; Katz, Connelly, and Wilson, 2007; Connelly and Katz, 2009) Katz and her colleagues have investigated the effect of reflection on learning in physics. In a series of studies, they gave students reflection questions after they had finished quantitative problem solving in the Andes (VanLehn et al., 2005) tutor. Typically these questions changed some aspect of the previously tutored problem, and asked the student to consider how the answer would change. After answering, the student would receive feedback in the form of either an interactive dialog or a “canned text” reading. Note that reflection was explicit in these studies: the student had to produce a visible answer to the reflection question. This is in contrast to “implicit” reflection which the student does internally, and which is not visible.

Reflection was found to improve the conceptual understanding of physics (Katz, Connelly, and Wilson, 2007). In addition, Katz, Allbritton, and Connelly (2003) found that the amount of *abstraction*, such as conceptual or strategic generalization, in a reflective dialog was correlated with learning. Finally, in several studies (Katz, Allbritton, and Connelly, 2003; Katz, Connelly, and Wilson, 2007) Katz and her colleagues found that the benefits of reflection could be gained using a "canned" text feedback, just as effectively as with interactive dialog feedback.

Together, these results show that explicit reflection can improve learning of physics after quantitative (not qualitative) physics tutoring. These results also suggest that the reflective intervention does not have to be implemented using interactive dialog, but that a text will be just as effective.

Informed by this work, the current study asked students to read a reflective text comparing similar parts of different problems which they had just completed in the Itspoke tutor. By comparing different applications of the same physics concepts, we hoped students would learn those concepts more deeply, and improve their post-test scores. We were also interested in the effectiveness of a reflective text because of the potential for a personalized text to be generated dynamically at run-time. We will return to this idea below.

A reflective text given after tutoring can only be effective to the extent in which it is processed and understood by the student. One important determinant of depth of processing has been shown to be reader interest. High interest readers have been found to form a deeper representation of a text's meaning than less interested readers, as indicated by measures of elaboration and accuracy in free recall (McDaniel et al., 2000; Schiefele and Krapp, 1996). Schiefele and Krapp (1996), for example, found that low interest readers built better verbatim representations of the text, but that high interest readers built better propositional representations. We therefore expect a role for interest in our implementation of reflective text. In particular, we gauge interest by measuring motivation to learn physics, and expect the reflective text to be less effective for poorly motivated students.

The literature therefore suggests two hypotheses for the current study. **Hypothesis One** is that a reflective/abstractive text will improve learning from the Itspoke tutor. **Hypothesis Two** is that the effect of this reading will vary with student motivation, with less impact on poorly motivated students.

Study Design

The Itspoke Tutor

Itspoke (**I**ntelligent **T**utoring **SPOKE**n dialog system) is a spoken dialog tutoring system which teaches five problems in qualitative physics. It was originally built by adding a spoken dialog interface to the Why2-Atlas (VanLehn et al., 2002) tutor, and recently re-implemented using the TuTalk (Jordan et al., 2007) dialog platform. At the start of each tutoring session, the tutor presents a problem statement and asks a question, to which the student responds. The tutor talks the student through the correct solution to this problem, entering sub-dialogs as necessary to remediate incorrect student answers. The version of Itspoke used in this study is the same as the automated version used in (Forbes-Riley and Litman, 2010), which does not require the student to write essays. Next we briefly describe two of the five qualitative physics problems taught by Itspoke.

In the **Earth-Sun** problem, the student is asked if the gravitational pull on the massive Sun from the lighter Earth is the same as the pull on the Earth from the Sun. The student is expected to realize that Earth and Sun form an action-reaction pair resulting from the gravitational pull between them. Newton's third law says that when a force is exerted by one body on a second body, the second body exerts a force of identical magnitude and opposite direction on the first body. Therefore the pull on each body will be the same.

In the **Car-Truck** problem, the student is asked to reason about the impact force between a lightweight car and a heavy truck, which are having a head-on collision. As in the Earth-Sun problem, the student is expected to recognize that car and truck form an action-reaction pair, and to reason that the impact force will be the same on each.

Experimental Conditions

Relevant aspects of our experimental design are shown in Table 1. All subjects read and signed a consent form, provided background information such as high school GPA and SAT scores, read a non-physics "warm-up" reading, then read introductory material about physics. After the introductory reading, they took a pre-test to measure their domain knowledge before tutoring.

Control **("read again")**	**Reflection** **("ref")**
Non-Physics Warmup Reading	Non-Physics Warmup Reading
Physics Pre-Reading	Physics Pre-Reading
Pre-Test	Pre-Test
Motivation Survey	Motivation Survey
Tutoring Dialogs	Tutoring Dialogs
Shortened Pre-reading	**Reflective Reading**
Post-test	Post-test

Table 1: Relevant aspects of the study design

Next, the subjects took a motivational survey, then engaged the Itspoke system in interactive tutorial dialogs which covered five physics problems. The Itspoke system was identical for all subjects. After tutoring, the subjects read a text which varied by condition. Subjects took an immediate post-test plus a delayed post-test after one week. [1]

For the post-tutoring reading in the control condition, subjects re-read a shortened version of the introductory physics text. In the reflection condition subjects read a reflective text which will be described more fully below. All readings were

[1] Due to space limitations, only our major findings are reported here. Results featuring far transfer and retention, extreme groups design, manipulations of the cohesiveness of the reflective text and measurements of cognitive load, will be included in other papers. The reader can also find them together in (Ward, 2010).

presented using the Linger interface (Rohde, 2003). The portion of the study which differed between conditions is shown in bold in Table 1.

We tested Hypothesis One, that adding a reflective reading improves learning, by comparing learning gains between the reflective and control ("read again") conditions. We expected that the reflective reading condition would show larger learning gains than the control. We tested Hypothesis Two by looking for an interaction between student motivation and the effect of reflection.

Participants

In total 166 students were recruited by flyer, by advertisement during an undergraduate psychology course, or from the University of Pittsburgh psychology subject pool. We adopted an "extreme groups" design (Feldt, 1961) to increase the power of a high-vs-low pre-test comparison which is reported elsewhere. [1] Therefore, 40 students whose scores fell in the middle third were dismissed after the pretest. An additional 27 were removed for incomplete data of various kinds, such as missing delayed post test scores.

This left a total corpus of 99 subjects. Subjects were assigned randomly to experimental conditions, which resulted in an allocation of 33 subjects to the control condition and 66 to the reflective reading condition. [2] 47 subjects were paid for their participation and the remainder were given credit toward their psychology subject pool requirement.

Motivational Survey

The motivational survey used in this study was adapted from work by Pintrich and Groot (1990), who developed the "Motivated Strategies for Learning Questionnaire (MSLQ)." The MSLQ includes questions which measure, among other things, the students' self-regulation behavior, attitudes about self-efficacy, and beliefs about the intrinsic value of the task. In this work we use a reduced version of the MSLQ, which is also patterned on an instrument used in previous work by Roll (2009)[3]. Our motivational survey is shown in Figure 1.

The dimensions of motivation measured are theoretically distinct. However, except for question three, responses to these questions were all very significantly correlated with each other in our survey ($p < .01$). We also found that our instrument's reliability improved when question three was omitted, as shown in Table 2.

Questions	Alpha
1, 2, 3, 4, 5	0.531
1, 2, 4, 5	0.716
2, 4, 5	0.703
4, 5	0.683

Table 2: Alpha

Please read the following statements and then click a number on the scale that best matches how true it is of you. 1 means "not at all true of me" whereas 7 means "very true of me".

1. I think that when the tutor is talking I will be thinking of other things and won't really listen to what is being said.
2. If I could take as much time as I want, I would spend a lot of time on physics tutoring sessions.
3. I think I am going to find the physics tutor activities difficult.
4. I think I will be able to use what I learn in the physics tutor sessions in my other classes.
5. I think that what I will learn in the physics tutor sessions is useful for me to know.

Figure 1: Pre-tutoring Motivational Survey

Table 2 shows values of Alpha (Cronbach, 1951) for various subsets of the motivation questions. We omit Question 3, which maximizes Alpha at .716. This is just above the commonly accepted (Gliem and Gliem, 2003; Cortina, 1993) threshold for reliability in such an instrument.

In order to look for an interaction between motivation and reflection, we divide students into "low," "middle" and "high" motivation categories based on our seven point motivation scale. Students whose average score was in the bottom half of the scale (below 3.5), are labeled "lowMot." Students who scored in the top half of the scale were given a median split, such that those who scored between 3.5 and 4.75 were labeled "midMot," and those who scored above 4.75 were labeled "hiMot." Although this method of categorization was primarily motivated by the semantics of the survey used, it divides students into fairly even thirds: 28 low, 36 middle and 35 high. However note from Table 5 that subjects are not distributed so evenly when further subdivided by motivational group.

Learning Measures

Our pre- and post-tests asked students to apply Newton's laws to situations not identical to the tutored problems. There were a total of 44 multiple choice questions on each test, with the post-test questions being isomorphic to the pretest questions. We report Normalized Learning Gain (NLG), which is computed as (post-pre)/(1-pre) where "post" is percentage correct on the post-test and "pre" is percentage correct on the pre-test (expressed between 0 and 1).

Designing the Abstractive/Reflective Texts

A major purpose of this work is to determine if a reflective/abstractive text can improve learning after qualitative physics tutoring. We now describe how the text was designed to perform these reflective and abstractive functions.

The reading compared places in which the same physics principle had been applied in different problem dialogs, and pointed out similarities in the overall problem solving approach between problems. It was structured to follow the four steps of reflection described by Tchetagni, Nkambou, and Bourdeau (2007) in the context of their reflective Prolog tutor. First, the readings elicited curiosity by asking about

[2] 32 students got a high cohesion, and 34 got a low cohesion version of the reflective text. These groups are combined in the "reflective" condition in the current analysis.

[3] We are very grateful to Maxine Eskenazi for providing the survey used in this study.

similarities between the tutored problems. They reviewed relevant parts of several of the problems; then pointed out which parts were common and essential, and which were unimportant. Finally, they demonstrated the correctness of the commonalities derived by showing how they would apply to another of the tutored problems.

For example, the two tutored problems described above both apply Newton's third law, but in different situations. Our reading attempts to help the student abstract their important commonalities by comparing these applications. In the Earth-Sun problem Newton's third law is used to motivate the idea that the earth pulls on the sun with exactly the same force as that with which the sun pulls on the earth (but in the opposite direction). In the car-truck problem, it is used to motivate the idea that the force of the car hitting the truck is the same as that of the truck hitting the car. The excerpts shown below demonstrate how Itspoke described these points to a student in two tutoring dialogs:

Dialog excerpts mentioning Newton's Third Law

From Earth-Sun: Okay. Newton's third law says every force has an equal and opposite reaction force. That is, if there is a force acting on object A due to object B, then there is also a force acting on B due to A. These two forces have the same magnitudes but opposite directions. Moreover, they are the same type of force. If one is a gravitational force, then so is the other. If one is a frictional force, then so is the other. In this case, there is a gravitational force on the earth due to the sun. Is there a gravitational force on the sun due to the earth?

From Car-Truck: Alright. Newton's third law says that every force has an equal and opposite reaction force. That is, if there is a force acting on object A due to object B, then there is also a force acting on B due to A. The two forces have the same magnitude and opposite directions. So in our problem, upon which vehicle is the impact force greater?

These two applications of Newton's Third Law were compared in the reflective reading, a portion of which is shown below. In step one, curiosity was elicited by reminding the student about certain parts of two problems, and asking about a similarity between them. This question is underlined in the excerpt shown. In step two, the relevant parts of the original problems were pointed out. In this case the fact that both used the Third Law is pointed out in the passage with double underlines. In step three, points of similarity and difference were mentioned. Passages performing this comparative function for two features of the Third Law are shown with a wavy underline. The first passage shows that in both problems, the type of force is the same for both interacting objects. The second passage shows that in both problems, the reaction force acts in the opposite direction to the action force. Other passages, not shown, point out unimportant surface differences between the problems. In step four, the commonalities were "evaluated" by applying them to a third problem. A portion of this segment is shown with a dotted underline.

Reflective Text Excerpt: Newton's Third Law

Newton's Third Law

In the Car-truck problem we wanted to compare the relative accelerations of the car and truck. Therefore, we first had to compare the impact force of the car on the truck with the impact force of the truck on the car. Similarly, in the Earth-Sun problem we were asked to compare the force of the Sun's pull on the Earth with the force of the Earth's pull on the Sun. Do you remember which of Newton's Laws was useful in these two problems?

In these two problems we used Newton's Third Law to show that the forces involved in an action/reaction pair had the same magnitude but acted in opposite directions to each other. An action-reaction pair is formed whenever one object exerts a force on a second object. Newton's Third Law says that when one object exerts a force on a second object, there is an equal and opposite reaction force from the second object back onto the first object. In addition, the type of force is always the same for both objects in the action/reaction pair. For example it was gravitational force on both Earth and Sun, and impact force on both car and truck. The two forces in an action-reaction pair can operate along any axis, but always have opposite directions to each other. For example, the earth pulled in the opposite direction than did the sun (vertically down vs vertically up), and the car's impact force was opposite to the truck's (horizontally right vs horizontally left).

...

For example in the Plane-Packet problem, the Earth exerts a gravitational force on the packet, and the packet accelerates downward toward the Earth. Does the Earth also accelerate toward the packet?

To the extent possible, each point in the reflective reading was constructed this way.

The read-again control text covered the same physics principles, but without explicitly referring back to the tutored problems. For example, below is an excerpt from its discussion of Newton's Third Law:

Read-again Control Excerpt: Newton's Third Law

...In the simplest sense, a force is a push or a pull. Looking closer, however, we find that a force is not a thing in itself, but is due to the interaction between one thing and another. One force is called the action force. The other is called the reaction force. It doesn‘t matter which force we call action and which we call reaction. The important thing is that neither force exists without the other. The action and reaction forces make up a pair of forces. ...

The read-again control had 2,013 words. It was a shortened version of the introductory reading, which had originally been developed for studies using the Why2 tutor. Some content was removed from this reading to control for length relative to the reflective reading. Sections were selected for removal if they covered concepts not in the reflective readings, with the result that both the post-test readings covered a similar set of physics topics. The two texts read by our reflective group, [2] had on average 1,851 words.

Results

Initial Results

To test if our reflective text improved learning we ran an Anova with normalized learning gain as the dependent variable and condition ("ref" vs "read-again") as the independent variable. Table 3 shows results for this Anova. The first column shows the p-value for the Anova. The second and third columns show the mean normalized learning gain, its standard deviation, and the N for the control and reflection groups, respectively. Note that while the mean learning gains favor reflection, the difference is not significant. This fails to support our first hypothesis on the entire data set.

	Mean NLG (SD) N	
pVal	**Read-Again**	**Reflective**
0.160	0.314 (.279) 33	0.379 (.172) 66

Table 3: Anova explaining NLG by reflection condition

Motivation Interaction Results

However, to test our second hypothesis, we next look for interactions between motivation and exposure to the reflective text. We test for interactions with motivation by running an Anova with normalized learning gain (NLG) as the dependent variable and both condition ("ref," "read again") and motivation category ("hiMot," "midMot," "lowMot") as the independent variables. Table 4 shows p-values for these two predictors and their interaction, with significant p-values in bold. The last column shows a significant interaction between experimental condition and motivation (F(2,93) = 4.23 p = .017). This suggests that the reflective reading has different effects on learning at different levels of motivation.

condition pVal	**motivation pVal**	**condition x motivation**
0.146	0.303	**0.017**

Table 4: Anova explaining NLG by reflection and motivation categories, and their interaction

Following the method described by Roberts and Russo (1999, p. 212), we next divided the corpus by level of motivation, to see how the effect of reflective text varied between levels. For each level of motivation, we ran an Anova with NLG as the dependent variable and reflective condition (cond) as the independent variable. Table 5 summarizes the results. The p-value for the Anova is in Column 2. Columns 3 and 4 show the mean NLG, its standard deviation, and the N for each condition. As shown in the middle row of Table 5, NLG for the middle motivation subjects is significantly higher in the reflective than in the non-reflective reading group (F(1,34) = 8.35 p = .007).

subj	**cond**	**Mean NLG (SD) N**	
Group	**pVal**	**Read-Again**	**Reflective**
lowMot	0.299	0.243 (.384) 6	0.360 (.190) 22
midMot	**0.007**	0.185 (.277) 11	0.410 (.182) 25
hiMot	0.222	0.429 (.194) 16	0.359 (.138) 19

Table 5: Anovas explaining NLG by reflection condition, for each motivation category.

These results support both our first and second hypotheses. They show that reading an abstractive/reflective text is better than reading the read-again control text for moderately motivated students. They also support our hypothesis that student motivation has important effects on the value of a reflective reading intervention.

Discussion and Related Work

In addition to the Katz studies described above, other researchers have investigated reflection in various domains. Several of these have also shown that the effect of reflection can vary with individual traits.

For example, Davis (2003) gave reflection prompts to students who were engaged in a "scientific evidence evaluation" task. Results suggested that students receiving nonspecific prompts to reflect learned significantly more than those receiving more specific prompts, and also that this effect was especially strong for students with high autonomy. Low autonomy students were thought to take less responsibility for their own learning, and so reflect less effectively when prompted. This may be similar to our result with less motivated students, suggesting that they reflected less effectively even when given a reflective text.

In a study by Lee and Hutchison (1998), students receiving reflection questions learned more than those in a no-reflection control, but this held only for students with low prior knowledge. Lee and Hutchison hypothesized that the high knowledge students were asking their own questions, and so did not need the reflection prompts in order to gain the benefits of reflection. This suggests an interpretation for our highly motivated students. Perhaps they also were reflecting even when not prompted, and so gaining the benefits of reflection from both types of text. Therefore only our middle motivation students, who were motivated enough to make use of the reflective text, but not so motivated as to reflect in all conditions, benefited from our intervention.

Other researchers have also found motivation to be important in tutoring. For example Graesser et al.(1995) found that a good human tutor will "bolster student motivation, confidence and self-efficacy" while learning.

Contributions and Future Work

By confirming Hypothesis one, this work suggests that reflection after *qualitative* physics tutoring can improve the learning of physics concepts, and that a reflective text is sufficient to produce this effect. While improved, however, learning was still fairly low. Among the middle-motivation students who were helped by our intervention, post test percentage correct in the reflective condition was .77(.1), while in the control condition it was .65(.14).

The success of this text, which focused on abstractive comparisons of previous problems, adds to evidence that abstraction has a role in reflection's effect on learning.

In confirming Hypothesis Two, this study has shown that student motivation is an important factor in the success of a reflective intervention. It further reinforces findings that models of non-cognitive student states such as motivation would be valuable additions to Intelligent Tutoring Systems.

Finally, the success of this intervention suggests that it could be valuable to dynamically generate personalized reflective texts after tutoring, determining their content by dialog performance. Decisions about when to present these texts could be informed by measurement of certain student affective states which may be related to motivation and interest. For example, work toward automatically detecting states such as uncertainty, flow and "zoning out" has already been done (Forbes-Riley and Litman, 2010; D'Mello and Graesser, 2006; Drummond and Litman, 2010). In future work, we hope to determine if dynamically generated reflective/abstractive texts can further improve learning, and if automatic state detection can help determine for which students such a text would be helpful.

Acknowledgments

This work was supported by the NSF (Grant # 0631930 and 0914615) as well as by the Learning Research and Development Center (LRDC) and the Intelligent Systems Program (ISP) at the University of Pittsburgh. We thank the Itspoke group, Alan Lesgold, Sandy Katz, Chris Schunn, Tessa Warren, Bob Hausmann and many others for helpful comments.

References

Connelly, J., and Katz, S. 2009. Toward more robust learning of physics via reflective dialogue extensions. In Siemens, G., and Fulford, C., eds., *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1946–1951. Chesapeake, VA: IOS Press.

Cortina, J. 1993. What is coefficient alpha? an examination of theory and applications. *Journal of Appl. Psych.* 78(1):98–104.

Cronbach, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3):297–334.

Davis, E. 2003. Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences* 2:91 – 142.

D'Mello, S. K., and Graesser, A. C. 2006. Affect detection from human-computer dialogue with an intelligent tutoring system. In *Intelligent Virtual Agents, 6th Intl. Conference, (IVA)*, 54–67.

Drummond, J., and Litman, D. J. 2010. In the zone: Towards detecting student zoning out using supervised machine learning. *Proceedings 10th International Conference on Intelligent Tutoring Systems (ITS).*

Feldt, L. 1961. The use of extreme groups to test for the presence of a relationship. *Psychometrika* 26(3):307–316.

Forbes-Riley, K., and Litman, D. J. 2010. Metacognition and learning in spoken dialogue computer tutoring. *Proceedings 10th International Conference on Intelligent Tutoring Systems (ITS).*

Gick, M., and Holyoak, K. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15:1 – 38.

Gliem, J., and Gliem, R. 2003. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. *Midwest Research to Practice in Adult, Continuing and Community Education*.

Graesser, A.; Chipman, P.; Haynes, B.; and Olney, A. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education* 48(4):612–618.

Graesser, A. C.; Person, N.; and Magliano, J. P. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology* 9:495 – 522.

Halloun, I., and Hestenes, D. 1985a. Common sense concepts about motion. *American Journal of Physics* 53(11).

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

Jordan, P.; Hall, B.; Ringenberg, M.; Cui, Y.; and Rosé, C. 2007. Tools for authoring a dialogue agent that participates in learning studies. In *Proc. of Artificial Intelligence in Ed., AIED*, 43–50.

Katz, S.; Allbritton, D.; and Connelly, J. 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.

Katz, S.; Connelly, J.; and Wilson, C. 2007. Out of the lab and into the classroom: An evaluation of reflective dialogue in andes. In *Proceeding of the 2007 conference on Artificial Intelligence in Education*, 425–432. Amsterdam, The Netherlands: IOS Press.

Lee, A., and Hutchison, L. 1998. Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied* 4:187–210.

Litman, D., and Silliman, S. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics.*

McDaniel, M. A.; Waddill, P. J.; Finstad, K.; and Bourg, T. 2000. The effects of text-based interest on attention and recall. *The Journal of Educational Psychology* 92(3):492–502.

Pintrich, P., and Groot, E. D. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology* 82(1):33–40.

Roberts, M., and Russo, R. 1999. *A student's guide to Analysis of Variance*. 29 West 35th St., NY: Routledge.

Rohde, D. 2003. Linger: a flexible platform for language processing experiments. http://tedlab.mit.edu/ dr/Linger/.

Roll, I. 2009. *Structured Invention Tasks to Prepare Students for Future Learning: Means, Mechanisms, and Cognitive Processes.* Doctor of philosophy, Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh, Pa.

Rosé, C. P.; Jordan, P.; Ringenberg, M.; Siler, S.; Vanlehn, K.; and Weinstein, A. 2001. Interactive conceptual tutoring in atlas-andes. In *Proceedings of AI in Education*, 256–266.

Schiefele, U., and Krapp, A. 1996. Topic interest and free recall of expository text. *Learning and Ind. Differences* 8(2):141–160.

Tchetagni, J.; Nkambou, R.; and Bourdeau, J. 2007. Explicit reflection in prolog-tutor. *International Journal of Artificial Intelligence in Education (IJAIED)* 17:169–215.

VanLehn, K.; Jordan, P. W.; Rosé, C. P.; Bhembe, D.; Boettner, M.; Gaydos, A.; Makatchev, M.; Pappuswamy, U.; Ringenberg, M.; Roque, A.; Siler, S.; and Srivastava, R. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proc. 6th Int. Conf. on Intelligent Tutoring Systems*, volume 2363 of *LNCS*, 158–167. Springer.

VanLehn, K.; Lynch, C.; Schulze, K.; Shapiro, J.; Shelby, R.; Taylor, L.; Treacy, D.; Weinstein, A.; and Wintersgill, M. 2005. The andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education* 15:147 – 204.

Ward, A. 2010. *Reflection and Learning Robustness in a Natural Language Conceptual Physics Tutoring System.* Doctor of philosophy, University of Pittsburgh, Pittsburgh, PA. 15260.