Interactivity and Expectation: Eliciting Learning Oriented Behavior with Tutorial Dialogue Systems

Carolyn Penstein Rosé & Cristen Torrey

Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA, 15260
cprose,ctorrey@cs.cmu.edu

Abstract. We investigate the reasons behind students' different responses to human versus machine tutors and explore possible solutions that will motivate students to offer more elaborated responses to computerized tutoring systems, and ultimately behave in a more "learning oriented" manner. We focus upon two sets of variables, one surrounding the students' perceptions of tutor qualities and the other surrounding the conversational dynamics of the dialogues themselves. We offer recommendations based on our empirical investigations.

1 Introduction

Recent classroom and laboratory evaluations of a wide range of learning technologies have revealed a disturbing phenomenon of unproductive student behavior [4,11,17,15] where students approach their interactions with them in a "performance oriented" manner, i.e., resorting to shallow strategies for getting through material as quickly as possible, rather than a "learning oriented" manner, i.e., trying to learn as much as possible. In this paper we explore the extent to which these patterns may be the result of a combination of a priori expectations and attitudes about the technology and features of the technology that enable students to engage in performance oriented behavior. We focus our investigations on tutorial dialogue systems [2,22,14,3,13,12]. A tutorial dialogue system is a type of state-of-the-art learning technology modeled after one-on-one human tutoring that engages students in natural language dialogues.

In tutorial dialogue interactions, the distinction between learning oriented and performance oriented behavior can be characterized in terms of patterns of student verbal behavior. For example, it may be manifest in terms of extremely terse, and sometimes non-existent, student responses to tutor questions, and an almost total lack of elaboration. Comparing student verbal behavior in response to humans and to tutorial dialogue systems both employing an equivalent typed chat interface, it was observed that students do not spontaneously offer the kinds of self-explanations they freely offer to human tutors when responding to equivalent questions from a computer tutor [21,22]. This poverty of self-explanation has a detrimental effect both on the tutoring system's ability to create an accurate model of student understanding and on the student's ability to master the material.

In this paper, we investigate the reasons behind students' different responses to human versus machine tutors and explore possible solutions that will motivate students to offer more elaborated responses to tutoring systems, and ultimately behave in a more "learning oriented" manner. Effecting a change in student behavior is one important step along the path towards increasing the effectiveness of tutorial dialogue technology. We focus upon two sets of variables, one surrounding the students' perceptions of tutor qualities and the other surrounding the conversational dynamics of the dialogues themselves. Our hypothesis is that the elaborations students freely offer to human tutors are motivated by interpersonal factors and by the interactive nature of dialogue with the tutor. Neither is commonly part of one's experience with computers. Thus, we hypothesize that we can induce students to generate more elaborated verbal responses generally, and self-explanations in particular, by elevating their a priori perceptions of the computer-based tutoring systems, by making the tutors more responsive and interactive, and especially by a combination of these two strategies.

2 Tutorial Dialogue Technology

We conduct our investigations of tutorial dialogue systems using a popular framework originally developed at the University of Pittsburgh, called Knowledge Construction Dialogues (KCDs) [22]. KCDs were motivated by the idea of Socratic tutoring, a highly interactive tutoring style evaluated favorably in comparison to a less interactive didactic tutoring style in [23]. KCDs are interactive directed lines of reasoning that are each designed to lead students to learn as independently as possible one or a small number of concepts, thus implementing a preference for an "Ask, don't tell" strategy. When a question is presented to a student, the student types a response in a text box in natural language. If the student enters a wrong or empty response, the system will engage the student in a remediation subdialogue designed to lead the student to the right answer to the corresponding question. Once the remediation is complete, the KCD returns to the next question in the directed line of reasoning.

3 Learning Oriented Versus Performance Oriented Behavior in Tutorial Dialogue Systems

Explanation is one of the key learning oriented behaviors students may engage in in a tutorial dialogue context. From a scientific viewpoint, one of the best substantiated educational findings in cognitive science research is the educational benefit of explanation, and in particular, the self-explanation effect [8,19]. Self-explanation benefits learners by revealing knowledge gaps, abstracting problem specific knowledge into schemas that can be applied to other relevant cases, and elaborating the representation of knowledge in the learners mind so that it can be more easily retrieved [26]. The self-explanation effect appears to be related to the process of constructing an explanation. Previous studies of human tutoring have revealed a significant correlation between amount of student explanation and learning [21,10].

Self-explanation has been frequently studied in connection with studies of the educational benefit of studying worked out example problems for mathematics and other problem solving domains. When students possess sufficient background knowledge and are sufficiently engaged, presenting them with correctly worked example problems in math and science and directing them to "self-explain" has been proven highly effective, even more effective than problem solving at early stages of skill acquisition in the context of laboratory studies [19,16]. Nevertheless, in classroom settings neither the appropriate level of background knowledge nor the ideal level of engagement with the material can be assumed. Thus, an important question for improving the state of education is how to design interactions with instructional technology that are effective for keeping students engaged and for supporting productive explanation activities in a way that would be practical to place in a classroom setting.

Explanation in a tutorial dialogue context is also important from an assessment standpoint. Previous studies of student-tutor interactions in a human tutoring setting have demonstrated a strong correlation between length of student response and likelihood for negative feedback offered from the tutor [21]. Thus, elaborate student explanations create more opportunities for valuable instruction by revealing knowledge gaps that might not otherwise come up. Increased awareness of student knowledge gaps facilitates the tutor's process of effectively adapting instruction to the individual needs of students. It also increases the likelihood that students notice their knowledge gaps and strive for deeper understanding [26].

Human tutors are highly successful at eliciting elaborated explanations from students [21,10] and highly successful at educating students [7,9]. In response to equivalent questions from a tutor, we have observed human tutors typically eliciting an order of magnitude more talk in a typed chat environment, including verbal self-explanation, from learners than a tutorial dialogue system in the same domain using similar prompts with an identical text-based chat interface. See Figure 1 for a typical interaction from the WHY-Atlas physics explanation tutor [25]. Notice the student offering a typical, one-word reply. Figure 2 presents an analogous interaction in a human tutoring scenario in the same domain using an equivalent typed chat interface. Not only does the student answer in a complete sentence, but, more importantly, the student also offers a justification for the answer. Since both of these interactions are typed, rather than in speech, the difference between student behavior with a human tutor and with a tutorial dialogue system cannot be explained as a modality difference.

In this paper we describe two studies, each of which explores an alternative explanation for this phenomenon in an attempt to understand better the reasons for unproductive student behavior with tutorial dialogue systems and to formulate a recommendation for a solution. First we explore the issue of the differences in student expectations of human tutors and of instructional technology. As a starting place, one potential explanation for the difference in behavior in response to human tutors and to tutorial dialogue systems is that the same norms of cooperativeness and politeness that strongly influence dialogue behavior in human-human discourse do not routinely apply in human-machine discourse. In addition to frequent one or two word answers in response to tutor questions in a tutorial dialogue system context, we have also observed students offering sarcastic comments about the system rather than answers

or sometimes entirely neglecting to answer tutor questions when they figure out that the system will continue to offer instruction even in the total absence of student effort to offer an answer. In rare cases, students proceed in this fashion for an entire dialogue.

Tutor: In order for your hand to feel pain from the impact, there must be a force acting on it. What force is acting on your hand?
Student: wall

Figure 1: This example illustrates a typical typed KCD interaction.

Tutor: There is need for some clarification. A body's motion is determined by the forces acting on it. So, what are the forces acting on earth?
Student: Since space has no gravity, then the only force acting on the earth is the pull of the sun.

Figure 2: This example illustrates a typical human tutoring interaction.

Thinking about the issue of a priori expectations more broadly, some previously published evidence already supports the position that the perception of computers as different from humans is a key factor leading to lack of explanation with tutorial dialogue systems specifically, and perhaps "performance oriented" behavior with computer tutors in general. Whereas a series of human-computer interaction studies by Reeves and Nass (2002) suggests that on some level people subconsciously treat computers like people, others have found that humans speak differently when they believe they are speaking to a computer rather than to a human, even when their partner uses identical language with them [24]. Schechtman and Horowitz (2003) focused on social issues such as politeness rituals, and not learning oriented behavior such as explanation. A similar recently published comparison between student verbal behavior with human tutors and with computer tutors shows that students not only display more politeness indicators in their natural language contributions to human tutors, but more "hedges" as well, perhaps as a face saving device [6]. Nevertheless, none of these previous studies focus on the specific issue of student explanations, although the specific issue of eliciting deep explanation behavior from students is particularly important for designing effective tutorial dialogue environments. We hypothesize that students will offer more explanation to an agent they believe is a human because of differences in expectations students bring with them about how they typically interact with humans versus how they typically interact with computers. Note that we are not attempting to overturn "The Media Equation" [18]. We are addressing HCI issues that affect the extent to which students engage in productive behavior for learning with instructional technology. Here we are simply arguing that while previous studies touch on similar issues, they do not specifically address this question, which is an important initial step for improving the effectiveness of instructional technology, particularly tutorial dialogue technology. Thus, our first study explores the impact of student expectations on explanation behavior.

From a different angle, we explore the contribution of limitations in the capabilities of the technology as a contributing factor to the problem. An alternative hy-

pothesis along these lines is that students will offer more explanation to a tutor agent that is more interactive because it will be perceived as more interested in their thoughts. One can easily hypothesize, for example, that the reason why students behave differently with tutorial dialogue agents than with human tutors is simply because the technology is still too rigid to engage in realistically natural dialogue interactions. Focused feedback is one important aspect of human tutorial discourse that sets it apart from tutorial dialogue agents. Human tutors exhibit a high degree of responsiveness to students. In contrast to human tutoring dialogues, no current tutorial dialogue systems are capable of acknowledging and offering tailored feedback for extended explanations that do more than answer a direct question asked by the tutor [2,22,12]. Focused feedback is one way that human tutors demonstrate to students that they are listening and understanding what the student is saying. Previous studies have substantiated the benefits of tutor feedback in assisting students in problem-solving tasks [5]. For this reason, our second study explores the impact of focused feedback on student explanation behavior.

4 Experimental Setup

The two studies reported in this paper shared many common experimental setup features, which we will describe in this section. Features that are specific to a single study will be described in the section below related to the specific study.

In both studies, students interacted independently with a tutor agent through a text-based chat interface at a computer terminal in a small student lab space. The chat setup can be configured in three different ways. In one mode, the student receives only automatically generated messages, produced by the KCD engine. In another, a human can edit each automatically generated message before sending it to the student. In a third mode, a human can compose the entire message. All three modes appeared identical to the student. Both the student and the tutor were able to view the history of the conversation in a scrolling dialogue history window at the top of the chat interface. A separate text input window was used for entering a text, and in the case of the tutor, entering and/or modifying an automatically generated text, before it was submitted.

The tutoring domain was basic college-level Newtonian physics, a domain in which the first author has researched the relative effectiveness of alternative instructional technologies for the past five years [20,21,22]. In both studies, the instructional manipulation was short, consisting of exactly one KCD dialogue designed to teach the concept of normal force, which is the force that every hard surface exerts on any object resting on its surface. As is common practice for tutoring studies, learning was assessed using a pre and post test. We used the same test for pre and post-test, which consisted of 5 multiple choice conceptual physics questions related to the concept of normal force covering all major points raised in the dialogue on normal force that is part of the experimental manipulation.

5 The Impact of Expectations Related to Humanness

In the first study we measured the impact of student expectations on student explanation behavior by comparing students interacting with a computer agent in two conditions. 40 university students participated in the experiment, one at a time, randomly assigned to each of the two conditions. In the experimental condition, students were told that they would be chatting with a human tutor. For the initial segment of their interaction through the chat interface, they conversed freely with a human about their extra-curricular interests. The purpose of this social interaction was to reinforce the idea that they were talking to a human. After several turns, the human shifted the chat mode to automatic tutoring using the KCD engine so that the topic shifted to the dialogue about normal force, and the tutor turns were generated completely automatically. The human remained in the loop just to introduce a delay in order to maintain the illusion that the student was still interacting with a human. In the control condition, the students were told that they would be chatting with a computer agent. There was no initial conversation about extra-curricular interests. And there was no delay introduced between when the student entered a conversational contribution and the tutor's automatically generated response was delivered. Note that there were two differences between the experimental and control conditions in this study. For our purposes it is not important to disentangle the source of elevated student expectation. The important point was to measure the impact that expectations about a tutor agent formed before a tutoring interaction affect student's behavior within that interaction.

After signing a consent form, each student took the pretest. The students then read a set of instructions that explained that they would be engaged in a dialogue about physics with a tutor agent. The only difference in the instructions between the two conditions was that in the experimental condition the students were told they would be interacting with a human tutor and in the control condition the students were told they would be interacting with a computer tutor. After the dialogue, the students took the post-test. After the post-test, the students filled out a questionnaire assessing the students' perceptions of the tutor agent as a manipulation check. One-on-one interviews with students after filling out the questionnaire served as an additional manipulation check. 90% of the students participating in the study believed what they were told about the tutor agent being either computer or machine. During this time the deception was also explained to the students, and they had the opportunity to make comments or ask any desired questions.

5.1 Contrasting Dialogue Behavior in Experimental and Control Conditions

While student verbal behavior in the control condition (see Figure 3) was similar to that observed in previous evaluations of KCDs [20], behavior in the experimental condition (see Figure 4) was strikingly different and much more like what had been observed in human tutoring corpora.

Notice in Figure 3, the control condition student answers two of the tutor's questions with one word answers. The third question, which is a why question, elicits a longer response as expected. Nevertheless, even the longer student response does not

offer much discussion. In contrast, in Figure 4, we see a student in the experimental condition answering the same three questions. Notice how much additional discussion the student offers in response to the second tutor question.

Two separate factors distinguished the control condition from the experimental condition and may have separately or jointly contributed to elevated student expectation in the experimental condition. First, the instructions explained that students would be talking to a human. Secondly, the initial portion of the chat discussion that was about their extra-curricular interests was highly interactive, showing great interest in the details of student interests. As mentioned above, for our purposes it is not important to disentangle the source of elevated student expectation. See Figure 5 for an example of one of these social interactions. Note that this initial conversation was completely devoid of physics content and could not contribute to the students' learning nor demonstrate any specific knowledge of physics on the part of the tutor.

Tutor: Ok, …So think about this: when a rock is falling, what force is acting upon the rock to make it fall?

Student: gravity

Tutor: Good. If a rock is resting on a table, is there still a gravitational force acting on the rock?

Student: Yes

Tutor: Correct. If gravity is pulling the rock down, why does the rock not fall through the table?

Student: Because the table is exerting a force as well

Tutor: Hard surfaces can support objects because they exert a normal force on the objects. That normal force opposes the gravitational force …

Figure 3. This example illustrates a typical interaction from Control Condition

Tutor: Ok,…So think about this: when a rock is falling, what force is acting upon the rock to make it fall?

Student: Gravity.

Tutor: Good. If a rock is resting on a table, is there still a gravitational force acting on the rock?

Student: Well, the force is present but the rock is not moving. I used to think of acting as requiring movement, so I am confused about this. I would think in this case the answer is supposed to be yes.

Tutor: Gravitational force exists everywhere on earth. If gravity is pulling the rock down, why does the rock not fall through the table?

Student: An equal and opposite force is acting on the rock.

Tutor: Hard surfaces can support objects because they exert a normal force on the objects. That normal force opposes the gravitational force…

Figure 4. This example illustrates a typical interaction from then Experimental Condition

We compared average student turn length between conditions focusing only on the portion of the dialogue interactions that were managed by the tutorial dialogue system

so that the stimulus was the same between conditions and only the expectation set up prior to the physics interaction influenced the difference in students' verbal behavior.

As mentioned, average student turn length in the control condition was consistent with that in previous evaluations of KCDs (mean = 3.6, standard deviation = 2.3). In the experimental condition, average turn length was much larger. Although it was not as large as previously reported average turn lengths of between 10 and 15 with human tutors in the same domain [20], it was significantly longer than in the control condition (mean = 8.6, standard deviation = 4.2, $t(38) = 4.78$, $p = 0.000$, effect size computed with Cohen's $d=1.49$). Thus, we observed a strong main effect of condition on raw amount of student talk per turn. Note that all t-test presented in this paper are unmatched 2-tailed t-tests except where otherwise noted.

Tutor: So before we get started with the physics part of the experiment, I would like to find out more about you. What do you like to do on the weekends?

Student: I only like to work on Artificial Intelligence. I am rather unusual in this regard. I spend all my time working on software.

Tutor: What kind of software are you working on right now?

Student: Well, I have written software that makes Debian packages semi-automatically out of large quantities of online software which is locates mainly using spidering and information extraction.

Tutor: Do you have a dream job? If so, what would it be?

Student: I would enjoy working at the higher end of military AI systems where I have access to all the tools and resources and can be isolated and develop so-called autistic states of consciousness.

Tutor: Sounds interesting! Ok, let's talk some physics now…

Figure 5. This example illustrates a typical social interaction

We decomposed student turns into idea units at clause boundaries in order to take inventory of how much additional information was communicated in the experimental condition. In the experimental condition, students uttered on average 1.86 idea units of elaboration per turn beyond the direct answer to the tutor's question (standard deviation = 1.93). In the control condition, students uttered only .5 idea units of elaboration per turn (standard deviation = .89). The difference was statistically significant ($t(38) = 2.87$, $p < .05$, effect size = .9 standard deviations).

5.2 Learning Gains Analysis

Our learning gains analysis provides some limited evidence that simply eliciting more explanation without any change in the actual interaction with students in the experimental condition yielded an increase in instructional effectiveness of the KCD technology. On average, students in both conditions knew about the same amount about the concept of normal force prior to their interaction with the instructional manipulation. Out of 11 possible points, students in the control condition earned a mean score of 7.8 on the pre-test, with a standard deviation of 2.44. Mean pre-test score in the

experimental condition was lower, although it did not differ significantly from this (mean = 6.9, standard deviation = 1.51). There was a significant main effect of test phase on learning over the whole population. Mean pre-test score was 7.34 with standard deviation 2.04. Mean post-test score was 9.73 with standard deviation 1.48. $t(38) = 6.06$, $p < .05$. So although the instructional manipulation as well as the pre/post test was short, students learned a measurable amount of physics knowledge from their interaction with the system. Based on previous results demonstrating a significant correlation between average student turn length and learning gains, and based on the large effect of condition on average student turn length, we expected to see a significant improvement in instructional effectiveness between the experimental condition and the control condition. What we found was less conclusive. There was a marginal effect approaching significance in favor of the experimental condition on learning gains in terms of adjusted post-test score using a 1-tailed t-test (Mean(experimental) = .66, standard deviation = .33, Mean(control) = .47, standard deviation = .44, $t(39) = 1.55$, $p = .06$).

Because KCDs use very simple language understanding technology to process student input, automatically generated tutor responses were not always appropriate to the student's answers. However, we verified that occasional inappropriate KCD feedback did not lead to a significant detrimental effect on student learning. For each student we computed eight separate tallies indicating number of correct answers treated as correct, correct answers treated as incorrect, incorrect answers treated as correct, incorrect answers treated as incorrect, correct elaborations treated as correct, correct elaborations treated as incorrect, incorrect elaborations treated as correct, and incorrect elaborations treated as incorrect. Since the KCD treated every answer as completely correct or completely incorrect, we treated each idea unit that was part of an answer as having been treated as correct or incorrect depending upon whether the answer to the question was classified as correct or incorrect by the KCD. The reliability of the human judgment for correctness versus incorrectness of idea units computed using Cohen's Kappa was computed at .78, so these tallies can be treated as reliable. We did not find any significant correlation between percentage of idea units treated correctly and adjusted post-test score or raw post-test score with or without effect of pre-test score factored out, either within or across conditions. Thus, we did not find any evidence that inappropriate feedback negatively impacted learning.

As in the KCDs used in [22], the KCD used in this study stepped students through a line of reasoning where students were lead through a series of applications of rules of physics in order to provide a foundation for an understanding of an individual conceptual rule of physics. Students answered questions about things they experienced in their every day lives to help them understand. For example, "If you hold a book in your hand, which way do you feel the book pushing?" Students can answer these questions even if they don't know any physics. They simply require students to think about their experiences. And yet, these questions help them to see laws of physics at work. The ultimate goal of a KCD is to bring students to a place where they can remember and articulate a single rule of physics. As part of that line of reasoning, students are eventually asked to go one step further and apply that rule. For example, after discussing the concept of normal force applied by a horizontally oriented object, students were asked to predict what would happen if the object was

now tilted. If students were not able to make the conceptual leap, their understanding was scaffolded using a subdialogue, which is an embedded line of reasoning. Eventually, if the students were not able to apply the rule with help, the rule was applied for them. The focus of this work was to provide conceptual help when students displayed a gap in their understanding with faulty problem solving actions. In the study reported in this paper, students were able to answer the KCD question most of the time. In fact, overall, only 10% of direct answers to KCD questions were incorrect, with equal numbers in both conditions. Thus, based on answers to questions in the main line of reasoning of the KCD, little need of remedial instruction was indicated. This is an indication that students were able to follow the KCD's line of reasoning effectively. However, it might also be an indication that the material was too simple to observe a difference in instructional effectiveness due to our experimental manipulations.

6 Study 2: Raising Expectations Through Feedback

The results of Study 1 confirmed our hypothesis that student expectation was a major factor leading to a lack of explanation behavior in previous tutorial dialogue research. Simply elevating expectations without any change in the technology significantly impacted the amount of learning oriented student behavior. However, while the experimental manipulation was effective for testing our hypothesis, using deception to raise student expectations is not a viable solution in practice, and thus is not sufficient by itself to provide the HCI community with specific interface design recommendations. Furthermore, while the increase in student explanation in the first study yielded a marginal increase in learning, an ideal solution would produce a statistically reliable effect on learning.

We hypothesized that offering additional feedback to students in response to their explanation behavior would yield a larger impact on learning. Thus, in a follow-up study we directly tested a second hypotheses, namely that increased interactivity in the form of focused feedback would yield an increase in student explanation behavior. We predicted that the enhanced interactivity and closer coordination would raise student expectations about the technology and communicate to the student more interest in their thoughts. We predicted this would lead to a similar increase in learning oriented behavior to what we observed in the previous study.

6.1 Experimental Design of Study 2

In the experimental condition, students received targeted feedback in addition to typical KCD responses, whereas in the control condition students received only typical KCD responses. Using the same chat setup but in a mode that allowed a human in the loop to edit tutor turns before they were presented to students, a human inserted focused feedback at the beginning of each tutor turn in the experimental condition, wherever possible. Thus, only in the experimental condition the tutor turns would contain more explicit connections with the particulars of what students said.

During a pilot testing phase, we noticed that increased interactivity by itself was not effective, in large part because the students' default taciturn behavior did not offer the tutor many opportunities to offer feedback. Thus, we introduced a modeling phase in both the experimental and control conditions in which students spent 5 minutes prior to their interaction with the tutor agent viewing a screen capture video of a student interacting with the tutor agent on an unrelated physics topic (i.e., the concept of displacement) and offering productive, learning oriented behavior.

In order to rule out the possibility that a difference in learning gains we observe between conditions is the result of additional instruction offered in the experimental condition, we assigned students to matched pairs when we randomly assigned them to conditions. We tallied the list of idea units offered to each student in the experimental condition and offered the same instructional content to the student's matched pair student in the control condition formulated as a reminder at the end of the KCD. Thus, we controlled for information presentation between conditions. Note that students were randomly matched. Thus, paired students did not necessarily have the same instructional needs. This is an important point since one key distinction between feedback per se and additional instruction more generally is whether it is tailored to the specific needs of the student. 20 local university students and staff participated in the study, 10 in each condition. Thus, there were 10 matched pairs of students.

Table 1. Types of Focused Feedback Used in Study 2

Category	**Tag Line**
Missing Answer	"You have not answered the question about___"
CloseAnswer	"___ is close. I would say ___."
CorrectAnswer/No Elaboration	"You're right, but let's think about why that is correct." "You're right, but let's think about the implications."
CorrectElaboration	"That's a good point about ___."
CorrectAnswer/ Incorrect Elaboration	"You're answer is correct, but your reasoning is not correct. It's not true that ___. "
Wrong Answer	"That's not correct. It's not true that ___."

In order to ensure that students in the experimental condition were treated equally, we developed a set of tag lines to use with different categories of feedback (See Table 1). This allowed us to control for tone so that the tutor's feedback would have a consistent feel with the automatically generated KCD responses. Correct Answer and Correct Elaboration feedback was meant to affirm students for positive behavior and encourage them to elaborate. These two types of feedback did not contain any instruction since they were not directed at any specific deficiency in the student's contribution by definition. The other classes of feedback were all forms of negative feedback, thus each referring specifically to at least one specific piece of content.

6.2 Impact on Student Explanation Behavior

As in the previous study, we found an effect of condition on student explanation behavior. In particular, there was a reliable difference in terms of number of idea units per turn included in elaborations according to a 1-tailed t-test (mean(experimental) = 3.5, standard deviation = 3.5, mean(control) = 1, standard deviation = 1.6, $t(19) = 2.3$, $p < .05$, effect size = 1.0). Thus, there is evidence that focused feedback has an impact on how much students say. As a rough, informal comparison, the effect size for difference in number of idea units was consistent across studies. However, in terms of raw average student turn length, the difference was only a statistical trend according to a 1-tailed t-test (Mean(experimental) = 5.83, standard deviation = 3.8, Mean(control) = 4.45, standard deviation = 2.3, $t(19) = 1.4$, $p=.17$). Note that raw average student turn lengths in both conditions were in between the two extremes observed in the previous study.

Another unexpected result was that a linear regression analysis demonstrated a stronger connection between positive feedback and average turn length and that between negative feedback and turn length. There was no significant correlation between amount of negative feedback and average student turn length (N=10, R-squared=.04, t=.53, p=.55). However, there was a reliable correlation between amount of positive feedback and student turn length (N=10, R-squared=.54, t=3.07, $p < .05$). It could be hypothesized that students who were correct more often became more confident because of their success at answering the tutor's questions, and their success was more of a factor leading to their increased levels of explanation rather than the feedback itself. Thus we examined explanation behavior in the control condition to assess whether there was a correlation between number of correct student answers and average student turn length, but there was no significant correlation. This supports the interpretation that it was the feedback and not the number of correct responses that influenced how much students explained in the experimental condition.

6.3 Learning Gains Analysis

There was no difference in learning between conditions. Difference in adjusted post test scores was not significantly different even with a 1-tailed t-test (Mean(experimental) = .64, standard deviation = .38, Mean(control) = .75, standard deviation = .33, $t(19) = 1.2$, $p = .24$). There was, however, a significant correlation between amount of explanation and learning within the control condition, even with effect of pretest score factored out (N=10, R2=.45, t=2.6, $p < .05$).

We then examined more closely the substantive feedback offered to students in connection with deficiencies on their pre-tests and corresponding performance on their post-tests. Only 3 students in the experimental condition received any substantive negative feedback. For two of those students the substantive feedback was directed at topics that were not tested on the post-test. The other student showed a knowledge gain between pre-test and post-test on the relevant concept addressed by the tutor's feedback. Thus, further exploration on the topic of feedback is required.

7 Recommendations and Future Work

While the results in this paper do not offer the final solution for overcoming the problem of unproductive student behavior and dramatically improving the instructional effectiveness of the technology, these studies yield some new insights and promising directions for continued investigation. We have presented two studies that demonstrate that both expectation and interactivity have a measurable impact on the extent to which students engage in productive behavior with instructional technology. The strong impact of artificially elevated expectations on explanation behavior and weak impact on learning observed in the first study offers confirmation that exploring ways of elevating student expectation is promising for improving student behavior and potentially contributes to enhancing the effectiveness of instructional technology, although it is not sufficient in itself for yielding a significant impact on learning gains. While increasing the amount of targeted feedback offered to students is a more viable option in practice than artificially elevating expectations through deception, the impact on student behavior based on that manipulation was not quite as pronounced and yielded no effect on learning. However, since our analysis from the second study demonstrated that only the types of feedback offered in response to correct answers in this study correlated with average student turn length, we plan to continue to investigate the potential use of this form of feedback in connection with other types of answers in order to yield a stronger impact on behavior overall. Here we only explored the use of positive feedback in connection with correct answers, but it is possible that similar forms of feedback in connection with incorrect responses would encourage students to elaborate more when they were less certain.

Acknowledgements

This work was funded by NSF SGER REC-0411483 and ONR Cognitive and Neural Sciences Division, Grant number N000140410107.

References

1. Aleven, V., Ogan, A., Popescu, O., Torrey, C., & Koedinger, K. Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation. *Proc. ITS 2004,* Springer Verlag (2004), 443-454.
2. Aleven V., Koedinger, K. R., & Popescu, O. A Tutorial Dialogue System to Support Self-Explanation: Evaluation and Open Questions. *Proc. AI-ED 2003*, IOS Press (2003).
3. Ashley, K. D., Desai, R., & Levine, J. M. Teaching Case-Based Argumentation Concepts Using Dialectic Arguments vs. Didactic Explanations. *Proc. ITS 2002*, Springer Verlag (2002), 585-595.
4. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proc. CHI 2004*, ACM Press (2004), 383-390.

5. Bangert-Drowns, R., Kulik, C., Kulik, J., & Morgan, M. The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research 61*, 2(1991), 213-238.
6. Bhatt, K., Evens, M. & Argamon, S. Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions. *Proc. Cog. Sci. Soc. 2004*, Erlbaum (2004).
7. Bloom, B. S. The 2 Sigma Problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher 13*, 1984, 4-16.
8. Chi, M. T. H., de Leeuw, N., Chiu, M. H., LaVancher, C. Eliciting self-explanations improves understanding. *Cognitive Science 18*, 3(1984), 439-477.
9. Cohen, P. A., Kulik, J. A., & Kulik, C. C. Educational Outcomes of Tutoring: A Meta-analysis of Findings. *American Educational Research Journal 19*, (1982), 237-248.
10. Core, M. G., Moore, J. D., Zinn, C. The Role of Initiative in Tutorial Dialogue. *Proc. European ACL 2003.*
11. Davis, J. M., Leelawong, K., Belynne, K., Bodenheimer, R., Biswas, G., Vye, N., & Bransford, J. Intelligent User Interface Design for Teachable Agent Systems. *Proc. IUI 2003*, ACM Press (2003), 26-33.
12. Evens, M. and Michael, J., (in press). *One-on-One Tutoring by Humans and Machines*, Lawrence Erlbaum and Associates, Mahwah, NJ.
13. Graesser, A., VanLehn, K., the TRG, & the NLT. Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and Accomplished Human Tutors on Learning Gains for Qualitative Physics Problems and Explanations, LRDC Tech Report, University of Pittsburgh, 2002.
14. Graesser, A. C., Bowers, C. A., Hacker, D.J., & Person, N. K. An anatomy of naturalistic tutoring. In K. Hogan & M. Pressley (Eds.), *Scaffolding of instruction.* Brookline Books, 1998.
15. Hietala, P. & Niemirepo, T. The Competence of Learning Companion Agents. *International Journal of Artificial Intelligence in Education 9*, (1998), 178-192.
16. Hummel, H., and Nadolski, R. (2002). Cueing for schema construction: Designing problem-solving multimedia practicals. *Contemporary Educational Psychology 27*, 2(2002), 229-249.
17. Leelawong, K., Davis, J., Vye, N., Biswas, G. The Effects of Feedback in Supporting Learning by Teaching in a Teachable Agent Environment. *Proc.ICLS 2002.*
18. Reeves, B., & Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Place*, Cambridge University Press, 1996.
19. Renkl, A. Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction 12*, (2002), 529-556.
20. Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., & VanLehn, K. Exploring the Effectiveness of Knowledge Construction Dialogues. *Proc. AI-ED 2003*, IOS Press (2003).
21. Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R., VanLehn, K. The Role of Why Questions in Effective Human Tutoring. *Proc. AI-ED 2003*, IOS Press (2003).
22. Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. Interactive Conceptual Tutoring in Atlas-Andes. *Proc. AI-ED 2001*, IOS Press (2001), 256-266.
23. Rosé, C. P., Moore, J. D., VanLehn, K., Allbritton, D., A Comparative Evaluation of Socratic versus Didactic Tutoring. *Proc. Cog. Sci. Soc. 2001*, Erlbaum (2001).
24. Shechtman, N., & Horowitz, L. Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and People. *Proc. CHI 2003*, ACM Press (2003).
25. VanLehn, K., Jordan, P., Rosé, C. P., and The Natural Language Tutoring Group. The Architecture of Why2-Atlas: a coach for qualitative physics essay writing. *Proc. ITS 2002*, Springer Verlag (2002).
26. VanLehn, K., & Jones, R. M. What mediates the self-explanation effect? Knowledge gaps, schemas or analogies? *Proc. Cog. Sci. Soc. 1993*, Erlbaum (1993), 1034-1039.