

Evaluating the Effectiveness of Tutorial Dialogue Instruction in an Exploratory Learning Context

Rohit Kumar, Carolyn Rosé, Vincent Aleven, Ana Iglesias, Allen Robinson

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
{rohitk, cp3a, va0e, alr}@andrew.cmu.edu

Abstract. In this paper we evaluate the instructional effectiveness of tutorial dialogue agents in an exploratory learning setting. We hypothesize that the creative nature of an exploratory learning environment creates an opportunity for the benefits of tutorial dialogue to be more clearly evidenced than in previously published studies. In a previous study we showed an advantage for tutorial dialogue support in an exploratory learning environment where that support was administered by human tutors [9]. Here, using a similar experimental setup and materials, we evaluate the effectiveness of tutorial dialogue agents modeled after the human tutors from that study. The results from this study provide evidence of a significant learning benefit of the dialogue agents.

1 Introduction

In this paper we evaluate the instructional value of an implemented tutorial dialogue system integrated with an exploratory simulation-based learning environment. Tutorial dialogue has long been argued to hold a great potential for improving the effectiveness of instruction that can be offered by intelligent tutoring systems. This claim is largely based on evidence from famous studies of expert human tutoring, where it was demonstrated to beat classroom instruction by two standard deviations [2,3]. Dialogue offers the potential for eliciting a high degree of cognitive engagement from students and offers tutors a great deal of flexibility in adapting the presentation of material to the individual needs of students.

While tutorial dialogue holds the potential for many benefits for the student, it also comes with a cost in terms of both time and energy. Vanlehn et al. (2005) present a review of a series of experiments comparing human tutoring to non-interactive control conditions. Surprisingly, the advantages of human tutoring are not consistently demonstrated across studies. In order for the benefits of tutorial dialogue to be demonstrated conclusively, the benefits must outweigh the cost. One potential explanation for the inconsistency in the pattern of results from previous comparisons of human tutoring to non-interactive alternatives is that some studies were conducted in environments that did not take advantage of the potential benefits of dialogue to a great enough extent for the benefits experienced by students to clearly outweigh the

cost. Consequently, our work is motivated more by the question of where tutorial dialogue might have the greatest impact on learning rather than evaluating whether dialogue is always a more effective form of instruction than an alternative. Thus, the research goal of the CycleTalk project, which forms the context for the work presented in this paper, is to evaluate the benefits of tutorial dialogue in an exploratory learning context where we hypothesize that the creative nature of the task will create an environment in which the benefits of tutorial dialogue will be more clearly evidenced than in previously published comparisons.

2 Motivation

The study presented in this paper builds on results from a previous study in which the students performed the same task with the same simulation environment but interacted with a human tutor rather than a tutorial dialogue system [9]. In that study, we compared three goal level conditions: The first condition was a script based learning condition (S) in which students worked through a set of written instructions that explained step-by-step how to move through the simulation space. The second condition was a slightly more exploratory problem solving condition (PS) where students were given their instructions at a higher level in terms of problem solving goals and would need to use means-ends analysis to derive the set of low level steps required to satisfy those goals. However, they were provided with reference material that contained all of the same information about how to achieve those goals as the students in the S condition. Thus, for all practical purposes, the only difference between the instructional materials provided to the PS condition students and those provided to the S condition students was the insertion of some extra section divisions and the labeling of the section headers. Furthermore, they used an augmented version of the simulation environment that allowed them to request hints during their problem solving. In a final, more exploratory condition, rather than being presented with an exact ordering of problem solving goals, students were provided with the same set of goals but told that they were free to address those goals in whatever order served their instructional objectives best. They were to negotiate the ordering with a human tutor who was there to support them. Thus, we referred to this condition as the Negotiated Problem Solving Goals (NPSG) condition. Because the students were able to interact with a human tutor, they used the original version of the simulation software that did not include the help button that the PS students had access to. The students in the NPSG condition, which was the only condition with dialogue-based support, learned the most out of the three conditions. In particular, they learned significantly more than students in the PS condition ($p < .05$) and marginally more than the students in the S condition ($p < .1$).

We consider these experimental results to contribute to the line of research investigating the trade-offs between human tutoring and non-dialogue control conditions, although the experimental setup is different in important ways from that used in previous comparisons. Consider the following series of empirical investigations. First, an evaluation of the AutoTutor system, a tutorial dialogue system in the domain of computer literacy, showed an advantage over re-reading of a textbook of about 0.5

standard deviations [8]. The textbook re-reading condition itself was no better than a no-treatment control condition. Similarly, a recent evaluation of WHY-AutoTutor, a system based on the same architecture as the original AutoTutor but applied to the domain of qualitative physics, demonstrates a significant advantage of this system over a textbook reading control [6]. However, in a different experiment the learning results obtained with WHY-AutoTutor were no worse than a *human tutoring condition* and yet not better than those in a control condition in which students read targeted “mini-lessons,” short texts that covered the same content as that presented in the dialogue [5]. In [9] as discussed in the previous paragraph, again we evaluated the merits of human tutoring (the NPSG condition) in comparison to two non-dialogue control conditions (the S and PS conditions). But note that the setup was different in important ways. First, students in all conditions in our study were presented with informationally equivalent reading materials. Rather than replacing the reading materials as in [5], the role of the human tutors in our study was to help students navigate and understand the materials. Secondly, the reading materials were neither as brief nor targeted to the test as the “minilessons” used in [5] nor were they as extensive as a text-book. Thus, the key difference is that because decisions about how to navigate the materials were required, there was a potential benefit to be gained from support in this navigation from the negotiation with the tutor that would result in appropriate tailoring of the material.

The purpose of the study presented in this paper is to evaluate the first implementation of the NPSG approach to instructional support in a simulation-based learning environment.

3 The CycleTalk System

We are conducting our research in the domain of thermodynamics, using as a foundation the CyclePad articulate simulator [4]. CyclePad was developed with the intention of allowing students to engage in design activities earlier in their education than was possible previously. Our explorations of CyclePad use focus on design and optimization of thermodynamic cycles, specifically Rankine cycles. A thermodynamic cycle processes energy by transforming a working fluid within a system of networked components (condensers, turbines, pumps, and such). Power plants, engines, and refrigerators are all examples of thermodynamic cycles. Rankine cycles are a type of heat engine that forms the foundation for the design of the majority of steam based power plants that create the majority of the electricity used in the United States. There are three typical paradigms for design of Rankine cycles, namely the Simple Rankine Cycle, Rankine Cycle with Reheat, and Rankine Cycle with Regeneration. As students work with CyclePad on design and optimization of Rankine Cycles, they start with these basic ideas and combine them into novel designs.

We have constructed a cognitive task analysis describing how students use CyclePad to improve a design of a thermodynamic cycle [10]. Students begin by laying out the initial topology of a cycle using the widgets provided by CyclePad. For example, they may choose to construct the topology for a Simple Rankine cycle, which

consists of a heater, a turbine, a condenser, and a pump. Students must next set values for key parameters associated with each widget until the cycle's state is fully defined. At that point, the student can explore the relationships between cycle parameters by doing what are called sensitivity analyses, which allow the student to observe how a dependent variable's value varies as an independent variable's value is manipulated. Students may experiment with a number of alternative designs. Based on their experience they can plan strategies for constructing cycle designs with higher efficiency. Making adjustments to improve cycle efficiency is called optimization. As part of this optimization process, students may reflect upon their understanding of how thermodynamic cycles work.

As a foundation for a tutorial dialogue system, we constructed a tutoring system backbone to integrate with CyclePad. The purpose of this tutoring system backbone was to introduce the capability of tracing the student's path through their exploration through the simulation space as well as to provide the capability of offering students hints along the way in the style of model tracing tutors. We used a tool set called the Cognitive Tutor Authoring Tools (CTAT) [7,1] to develop this backbone tutor. The Cognitive Tutor Authoring Tools (CTAT) support the development of so-called Pseudo Tutors, which can be created without programming, namely, by demonstrating correct and incorrect solutions to tutor problems, which are then stored in a representation referred to as a Behavior Graph, which is then used to trace the solution paths students follow as they are working with the Pseudo Tutors at run time. Each node in the Behavior Graph represents an action a student may make. We integrated tutorial dialogue with the pseudotutors by allowing dialogues to be launched either in response to hint requests or as success messages when students complete the action associated with a node in the Behavior Graph. We build on the foundation of our prior work building and evaluating Knowledge Construction Dialogues (KCDs)[11].

4 Content Development for Tutorial Dialogues

Using an exploratory corpus analysis tool called InfoMagnets that was developed in our research group, a thermodynamics domain expert constructed a topic analysis of a corpus of human tutoring dialogues collected during the previous classroom study involving CyclePad mentioned earlier [9]. Altogether each student's protocol was divided into between 10 and 25 segments such that the entire corpus was divided into approximately 379 topic segments altogether. The resulting analysis consisted of 15 distinct topics, with each student covering between 4 and 11 of these topics either once or multiple times throughout their interaction with the tutor. In support of the instructional value of these discussion topics, we found a strong and reliable correlation between topic coverage (i.e., percentage of topics discussed at least once) and post-test score, with pre-test score used as a covariate ($R^2=.715$, $N=21$, $p < .05$). The topic analysis of the corpus gives us a way of quickly getting a sense of how tutors divided their instructional time between different topics of conversation. Based on our topic analysis of the human-tutoring corpus, we designed 12 KCDs, 7 of

which were focused on the specifics of optimization and 5 of which were more conceptual.

Here is an example collected during our study of a student conversing with a KCD about manipulating the maximum temperature of the cycle.

Tutor: Let's stop for a minute and reflect on what we have seen. Having seen how changing T-Max (i.e., the Maximum Temperature at the Heater Outlet) affects the efficiency of the cycle, do you think it is a good idea in general for the Maximum Temperature to be increased and kept high?

Student: *yes*

Tutor: Okay. Let's think now about one important consequence of keeping T-Max high. What happens to the steam quality at the turbine exit if you increase the maximum temperature of the cycle?

Student: *Increases*

Tutor: Right. As increasing the temperature makes the efficiency and steam quality better, we should increase the maximum temperature of the cycle. But can we keep on increasing the maximum temperature at the heater outlet indefinitely, or are there limitations to keep in mind?

Student: *there are limitations such as material constraints*

Tutor: Great. we should try to keep T-Max (i.e., the maximum temperature at the Heater outlet of the cycle) as high as possible without risking safety concerns or exceeding the maximum temperature the material can withstand. Keeping T-Max high increases efficiency and keeps the steam quality high, which are both important. This is an important principle to keep in mind while optimizing the efficiency of your cycle.

5 Method

The purpose of our study was to test whether we can achieve the same benefit with an implemented tutorial dialogue system as we obtained with the presence of a human tutor in the NPSG condition from [9].

Experimental procedure common to all conditions. The study consisted of a 3 hour lab session. We strictly controlled for time between conditions. The 3-hour lab session was divided into 9 segments: (1) After completing the consent form, students were given 15 minutes to work through an introductory exercise to familiarize themselves with the CyclePad software. (2) Students then had 15 minutes to work through a 50 point pre-test consisting of short answer and multiple choice questions covering basic concepts related to Rankine cycles, with a heavy emphasis on understanding dependencies between cycle parameters. (3) Students then spent 15 minutes reading an 11 page overview of basic concepts of Rankine cycles. (4) Next they spent 40 minutes working through the first of three focused materials covering the Basic Rankine Cycle. The materials included readings, suggested problem solving goals,

and analyses to help in meeting those goals. (5) Next they spent 20 minutes working through the second set of focused materials, this time focused on Rankine Cycles with Reheat. (6) They then spent 20 minutes through the third set of focused materials, this time focusing on Rankine Cycles with Regeneration. (7) They then spent 10 minutes on each of two Free Exploration exercises, one of which was designed to test whether students learned how to fully define a Rankine cycle, and one of which was designed to test the student's ability to optimize a fully defined cycle. (8) They then spent 20 minutes taking a post-test that was identical to the pretest. (9) Finally, they filled out the questionnaire. The experimental manipulation took place during steps (4)-(6).

Experimental design. Our experimental manipulation consisted of 3 conditions. The only difference between conditions during the experimental manipulation was the version of the software the students used. In the control condition, students used the original CyclePad system. This was a replication of the script condition (S) from [9]. In the first experimental condition, students used a version of CyclePad augmented with feedback and help that were integrated with CyclePad using psuedotutors (PSHELP). The PSHELP condition was similar to the problem solving condition (PS) from [9] except that in addition to typical hints and feedback messages, dialogues in the form of Knowledge Construction Dialogues (KCDs) were attached to nodes related to KCD topics in such a way that if a student asked for help on that node, they would get the dialogue as the hint message. Thus, students only saw dialogues when they asked for help on nodes that had KCDs attached to them. In a second experimental condition, the same KCDs were attached to success messages on the same nodes so that students got the dialogues after they successfully completed an action or if they asked for help on that action (PSSUCCESS). In both experimental conditions, students only viewed each unique KCD once. If additional opportunities to view the same KCD came up, students instead were presented with a hint summarizing the message of the KCD.

Outcome Measures: We evaluated the effect of our experimental manipulation on three outcome measures of instructional effectiveness. One outcome measure was assessed by means of a Pre/Post test containing 32 multiple choice and short answer questions that test analytical knowledge of Rankine cycles, including relationships between cycle parameters. A domain expert associated each question on the test with the set of concepts related to the 12 authored KCDs discussed in Section 4 that the student would need to have a grasp on in order to correctly answer the problem. Using this topic analysis of the test, we can compute a concept specific score for each student on each test, and thus measure concept specific knowledge gain. Next there were two separate measures of practical knowledge, based on success at the two Free Exploration exercises from step 7 of the experimental procedure. For the Free Exploration 1 exercise where students were charged with the task of fully defining a Rankine cycle, they received a 1 if they were successful and 0 otherwise. For Free Exploration 2, the students were evaluated on their ability to optimize a fully defined cycle. Thus, their score for that exercise was the efficiency they achieved, as measured by the CyclePad simulator.

Participants. 31 students from a sophomore Thermodynamics course at Carnegie Mellon University participated in the study in order to earn extra credit. The study took place one and a half weeks after Rankine cycles were introduced in the lecture portion of their class. The study took place over two days, with two lab sessions on each day.

6 Results

The goal of our evaluation was to measure the value added of dialogue to the Cycle-Talk system. Our two experimental conditions present two different approaches to integrating dialogues with a version of CyclePad that was augmented with an intelligent tutoring framework that allowed students to ask for hints. Students had the opportunity to encounter two different types of dialogues. In particular, 5 dialogues covered basic knowledge about the concept of Reheat, the concept of Regeneration, and some basic knowledge about properly initializing cycle parameters prior to optimization. 7 additional KCDs covered specific topics related to interpreting sensitivity analyses and doing optimization based on the results.

As mentioned, the difference between the PSHELP condition and the PSSUCCESS condition was that students in the PSHELP condition only received KCDs in response to help requests whereas students in the PSSUCCESS condition also received KCDs as success messages after successfully completing a sensitivity analysis. Thus, the PSSUCCESS condition included more paths where students had the opportunity to encounter KCDs, although the system ensured that each full KCD was never viewed by the same student more than once. Students in the PSSUCCESS condition were significantly more likely to see each KCD than students in the PSHELP condition, as computed from logfile data using a binary logistic regression with an observation for each KCD (i.e., whether the student viewed that KCD or not during their experience with CyclePad) for each student in the two experimental conditions ($p < .05$). Students in the PSHELP condition only viewed a KCD when they asked for a hint. And in practice, students in the two experimental conditions did not ask for help frequently. Specifically, only about 14% of the problem solving actions of students were help requests. On average, students in the PSHELP condition viewed 1.8 KCDs (st. dev. .837) whereas students in the PSSUCCESS condition saw 2.7 (st. dev. 1.9). The difference in coverage of KCDs between conditions was mainly on the KCDs related to sensitivity analyses. Only 1 of 7 KCDs focusing on interpreting sensitivity analyses was viewed by any student in the PSHELP condition, whereas in the PSSUCCESS condition all but one of these KCDs was viewed by at least one student. The difference between experimental conditions is interesting from the standpoint of evaluating the contribution of manipulating the number of KCDs viewed on learning. Nevertheless, it is a concern that so few of the authored KCDs were viewed by students on average even in the condition where they were viewed most frequently, and further increasing the number of opportunities for students to view KCDs is one of the goals of our continued work.

As mentioned above, the study took place over two days, with two lab sessions on each day. Two lab sessions on day 1 were assigned to the control condition (S). The first lab session on the second day was assigned to the first experimental condition (PSHELP). The final experimental condition took place during the second lab session on the second day (PSSUCCESS). We learned after the experiment was in progress that a quiz on Rankine cycles was administered to the class in between the lab sessions on the first day and the lab sessions on the second day. Thus, presumably because students were studying the day before the quiz, on average pretest scores increased from lab session to lab session such that there was a weak but significant correlation between lab session number and pretest score ($R\text{-squared} = .14, p < .05, N=17$). We expect that students on the second day when the experimental conditions took place were less motivated to learn the material than students on the first day since the quiz had already been given. Furthermore, since they had already studied, any learning that would take place would necessarily need to be on topics that remained difficult for the students even after studying. Finally, attendance in the first lab on the second day was lower than in the other conditions. Because of the interference of the quiz between the lab sessions where the control condition was conducted and the lab sessions where the two experimental conditions were conducted, we disregard the comparison between the control condition and the two experimental conditions and focus only on the difference between the two experimental conditions, although a summary of results from all three conditions is displayed in Table 2.

Table 1 Summary of results from all three conditions.

Condition	Pretest Average Total	Posttest Average Total	FreeExplore 1 Success Rate	FreeExplore 2 Average Efficiency
S	20.64 (5.56)	31.39 (5.86)	23%	38.14 (10.97)
PSHELP	20.67 (3.56)	27.83 (6.02)	0%	38.09 (13.12)
PSSUCCESS	24.86 (4.10)	32.45 (4.06)	20%	34.09 (14.17)

There was a significant effect of test phase $F(1,60) = 44.98, p < .001$, with no significant interaction with condition. Thus, students in all three conditions learned. It is impressive that the lab sessions on the second day when the experimental conditions were conducted yielded significant learning gains even though they were conducted on the same day as the quiz, which took place that morning. Because the difference in presentation of KCDs between the two experimental conditions is subtle, in order to increase the statistical power of the comparison, we evaluated the significance of the difference in learning between conditions using a repeated measures ANOVA, with a separate observation for each of the concepts the pre/post test was designed to test. The effect of condition on Concept Posttest Score with Concept Pretest Score used as a covariate and Concept as a fixed factor demonstrated a significant effect

both of Concept, $F(10, 327) = 15.55$, $p < .001$, and of Condition, $F(2, 327) = 3.25$, $p < .05$, with no significant interaction. Thus students learned more about some concepts than others consistently across conditions. A pairwise Tukey test comparing learning between the PSHELP and PSSUCCESS conditions demonstrated significantly more learning in the PSSUCCESS condition, which is the condition where more KCDs were viewed ($P < .05$), effect size .35 standard deviations. Thus, manipulating the number of KCDs viewed had a significant positive effect on student learning, although there were no significant effects of condition on either of the Free-Exploration exercises.

In a follow-up study with the same materials conducted at the US Naval Academy where we contrasted the S and PSSUCCESS conditions, we confirmed a significant effect in favor of the PSSUCCESS condition $F(1,86) = 5.57$, $p < .05$, effect size .25 standard deviations.

7 Discussion and conclusions

In this paper we presented results from a study that demonstrate the instructional effectiveness of Knowledge Construction Dialogues (KCDs) modelled after the human tutors from the study previously published in [9]. Nevertheless, the system we evaluated in this study still falls far short of a full implementation of the NPSG condition from that study. In our current work we are exploring ways to increase the similarity between our implemented tutorial dialogue system and the behavior of the human tutors from the NPSG condition. In particular, the number of KCDs students view during their experience with the system still need to be increased by a factor of 2 or 3 to bring it more in line with the number of topics covered in discussions with the human tutors from the human tutoring study. Furthermore, while the content development for the KCDs evaluated in this study were motivated by an analysis of the human tutoring corpus from the previous study, they played more of a role of eliciting reflection from students rather than assisting with navigation to the same extent that the human tutors did.

Acknowledgements

This project is supported by ONR Cognitive and Neural Sciences Division, Grant number N000140410107.

References

1. Alevin, V, & Rosé, C. P. (2005). Authoring plug-in tutor agents by demonstration: Rapid rapid tutor development, *Proceedings of AI in Education '05*.
2. Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16.

3. Cohen, P. A., Kulik, J. A., and Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19:237–248.
4. Forbus, K. D., Whalley, P. B., Evrett, J. O., Ureel, L., Brokowski, M., Baher, J., Kuehne, S. E. (1999). CyclePad: An Articulate Virtual Laboratory for Engineering Thermodynamics. *Artificial Intelligence* 114(1-2): 297-347.
5. Graesser, A., VanLehn, K., the TRG, & the NLT. (2002). *Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and Accomplished Human Tutors on Learning Gains for Qualitative Physics Problems and Explanations*, LRDC Tech Report, (2002) University of Pittsburgh.
6. Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., Chipman, P., Franceschetti, D., Hu, X., Louwerse, M. M., Person, N. K., and the Tutoring Research Group, (2003). Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog. *Proceedings of the Cognitive Science Society*.
7. Koedinger, K. R., Alevan, V., Heffernan, N., McLaren, B., & Hockenberry, M. (2004). Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. In J. C. Lester, R. M. Vicario, & F. Paraguaçu (Eds.), *Proceedings of Seventh International Conference on Intelligent Tutoring Systems, ITS 2004* (pp. 162-174). Berlin: Springer Verlag.
8. Person, N., Bautista, L., Graesser, A., Mathews, E., & The Tutoring Research Group (2001). In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future, Proceedings of AI-ED 2001* (pp. 286-293). Amsterdam, IOS Press.
9. Rosé, C. P., Alevan, V., Carey, R., Robinson, A., Wu, C. (2005). A First Evaluation of the Instructional Value of Negotiatble Problem Solving Goals on the Exploratory Learning Continuum, *Proceedings of AI in Eduction '05*
10. Rosé, C. P., Torrey, C., Alevan, V., Robinson, A., Wu, C. & Forbus, K. (2004). CycleTalk: Towards a Dialogue Agent that Guides Design with an Articulate Simulator, *Proceedings of the Intelligent Tutoring Systems Conference*.
11. Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., Weinstein, A. (2001). Interactive Conceptual Tutoring in Atlas-Andes, *Proceedings of AI in Education 2001*
12. VanLehn, K., Graesser, A., Tanner, J., Jordan, P., Olney, A. & Rosé, C. P. (2005). When is reading just as effective as one-on-one interactive tutoring? *Proceedings of the Annual Meeting of the Cognitive Science Society*