Evaluating a General Model of Adaptive Tutorial Dialogues

Amali Weerasinghe[1], Antonija Mitrovic[1], David Thomson[1],
Pavle Mogin[2], Brent Martin[1]

[1]Intelligent Computer Tutoring Group, University of Canterbury, New Zealand
[2]Victoria University of Wellington, Wellington New Zealand
{amali.weerasinghe, david.thomson}@pg.canterbury.ac.nz, pavle.mogin@ecs.vuw.ac.nz,
{tanja.mitrovic, brent.martin}@canterbury.ac.nz

Abstract: Tutorial dialogues are considered as one of the critical factors contributing to the effectiveness of human one-on-one tutoring. We discuss how we evaluated the effectiveness of a general model of adaptive tutorial dialogues in both an ill-defined and a well-defined task. The first study involved dialogues in database design, an ill-defined task. The control group participants received non-adaptive dialogues regardless of their knowledge level and explanation skills. The experimental group participants received adaptive dialogues that were customised based on their student models. The performance on pre- and post-tests indicate that the experimental group participants learned significantly more than their peers. The second study involved dialogues in data normalization, a well-defined task. The performance of the experimental group increased significantly between pre- and post-test, while the improvement of the control group was not significant. The studies show that the model is applicable to both ill- and well-defined tasks, and that they support learning effectively.

Keywords: adaptive tutorial dialogues, constraint-based tutors, Ill-defined tasks, well-defined tasks

1. Introduction

One of the aspirations of AIED research is to explore how intelligent systems can achieve the same effectiveness as in human one-on-one tutoring. One of the major factors contributing to the effectiveness of human tutors is the conversational aspect of instruction. Dialogues provide opportunities for students to reflect on their existing knowledge and to construct new knowledge. Some of the existing dialogue-based tutoring systems are Why2-Atlas [1], Auto Tutor [2], CIRCSIM-Tutor [3], Geometry Explanation Tutor [4] and KERMIT-SE [5]. Why2-Atlas and Auto Tutor use dialogues as the main learning activity, while the others provide problem-solving as the main activity and use tutorial dialogues as a way of remediating student errors. For example, CIRCSIM-Tutor is a natural language tutor that helps students learn cardiovascular physiology related to regulation of blood pressure. The Geometry Explanation Tutor requires students to justify the problem-solving steps in their own words. KERMIT-SE, a database design tutor, engages students in dialogues when their solutions are erroneous. All these tasks except database design are well-defined: problem solving is well-structured, and therefore explanations expected from learners can be clearly defined. In contrast, database design is an ill-defined task: the final result is defined only in abstract terms, and there is no algorithm to find it [6].

Our goal is to develop a general model for supporting dialogues across domains. Based on the findings of two Wizard-of-Oz studies [7], we developed a model consisting of three parts: an error hierarchy, tutorial dialogues and rules for adapting them. The error hierarchy categorizes all error types in a domain. At the leaf level, an error type is associated with one or more violated constraints. (The knowledge bases of our constraint-based tutors are represented in terms of constraints.) The error types are then grouped into higher-level categories. Remediation is facilitated through tutorial dialogues, one of which is developed for each error type. When there are multiple errors in a student solution, the hierarchy is traversed to select the error most suitable for discussion and the corresponding dialogue is then initiated. Finally, the adaptation rules are used to individualize the dialogues to suit the student's knowledge and reasoning skills by controlling their timing and the exact content. In response to the generated dialogue learners are able to provide answers by selecting an option from a list. For a detailed discussion of the model see [7].

In this paper we discuss how we evaluated the effectiveness of our model supporting an ill-defined and a well-defined task. The first study investigated the effectiveness of our model in database design (an ill-defined task), in the context of EER-Tutor [8]. In database design, students design database schemas using the EER model. Students need to know the concepts of the EER data model, use world knowledge about different real-world scenarios (i.e. enrolling students in a university etc.) and be able to handle the ill-definedness of the task. In the second study, we evaluated our model in data normalization, using NORMIT [8]. Data normalization is the process of refining a relational database schema in order to ensure that all relations are of high quality. This task requires normalizing a given database schema using the specified procedure. NORMIT contains a page for each step of this procedure, and students are requested to complete one step before continuing with the next one. The following two sections present the results of the study, followed by discussion and conclusions.

2. EER-Tutor Study

We conducted a study with the EER-Tutor in March 2010 at the University of Canterbury, which involved volunteers from an introductory database course. The objective of the study was to investigate whether adaptive dialogues are more effective in improving learning than non-adaptive dialogues in database design.

The participants were randomly assigned to groups. The experimental group received adaptive dialogues, while the control group had non-adaptive dialogues. The differences between the two groups were in dialogue selection, dialogue prompts and additional support. Dialogues for the control group were selected using the depth-first traversal of the error hierarchy. The first violated constraint that was found in the traversal was selected for discussion. As the errors in the hierarchy are ordered from simpler to more complicated errors, the depth-first search results in the simplest error for the control group.

The dialogues in our model consist of four stages [7]: (i) a problem-independent prompt discusses the relevant domain concept for the selected error; (ii) a problem-dependent prompt discusses the error in the context of the current problem; (iii) a corrective action prompt provides an opportunity to understand how to correct the error and (iv) a reinforcement prompt, providing another opportunity to learn the related domain concept. The control group saw the entire dialogue regardless of the number of times they have seen the dialogue previously or their responses to the dialogue prompts. As the result, the same solution submitted by two different students with different knowledge levels in the control group received identical dialogues. In contrast, an experimental group participant receives the problem-dependent prompt (prompt (ii)) the first time a mistake is done. If s/he makes this type of error repeatedly, the dialogue will start from the problem-independent prompt. The exit point of the dialogue for the experimental group is customized based on the student's past interactions with the dialogues. For a detailed description, see [7].

When an experimental group participant abandons a problem (i.e. changes a problem without attempting it) or has been inactive for a period of time, they were asked whether they needed help. If they requested help then their solution was evaluated and an error was selected for discussion based on their student model. The control group did not receive this support.

The study consisted of four stages: pre-test, interactions with EER-Tutor, post-test and questionnaire. The pre- and post-tests had 6 questions each, of similar difficulty. We wanted to evaluate whether students' problem-solving abilities as well as explanation skills improved after interacting with the system. One question asked the participants to provide the database schema for the given requirements. This is a typical question that can be found in examinations, text books etc. The other three questions were aimed to understand the effect the system had on students' explanation skills.

The participants used EER-Tutor for the first time in their regular lab sessions during the third week of the course, for a single 2-hour session. At the beginning of the session students were given about 10 minutes to complete the pre-test, after which they interacted with the system. Towards the end of the session, they were given 10 minutes to complete the post-test and 5 minutes to answer a questionnaire.

Out of 104 students enrolled in the course, 77 participated in the study. There was no significant difference in the pre-test performance between the control and the experimental groups. Some students have not completed the post-test. Table 1 reports some statistics about the 65 participants who completed both pre-and post-tests.

Table 1. Some statistics from the EER-Tutor study (sd given in parentheses)

	Control (34)	Experimental (31)	p
Pre-test (%)	54.5 (18.1)	51.3 (16.1)	ns
Post-test mean (%)	61.2 (14.9)	69.9 (11.5)	0.005
Gain	6.8(15.6)	18.6 (16.8)	0.002
Normalised gain	0.002 (0.7)	0.3 (0.4)	0.01
Interaction time (min)	62.8 (22.1)	62.9 (24.1)	ns
Attempted Problems	8.6(4.8)	10.6(4.8)	ns
Solved problems	9.0(4.8)	7.9 (4.7)	ns
Total Dialogues received	12.1 (7.3)	14.0 (8.3)	ns
Questions answered	34.4 (25)	23.6 (14.6)	0.01
% of correct answers	61.4 (23.1)	59 (16.9)	ns

There were 31 participants in the experimental group and 34 in the control group, with no significant difference on the pre-test performances. The post-test performance of the experimental group was significantly better compared to their peers who received non-adaptive dialogues. Both the learning gain (post-test score – pre-test score) and the normalised learning gain[1] of the group who received adaptive dialogues was also significantly higher than the gains of the control group.

There were no differences between the times spent with the system, the numbers of attempted and solved problems, and the number of dialogues received. The control group answered a significantly higher number of questions than their peers. This was expected, as the control group had to go through the entire dialogue before resuming problem-solving. However, percentages of correct answers are similar for both groups.

The effect size[2] (Cohen's d) for learning gains of the two groups is 0.69 (the effect size based on the normalized gain is 0.51). The effect size obtained here is remarkable because the only difference between the two groups was the adaptivity of the dialogues.

In order to investigate how the students learnt the database design concepts in terms of constraints, we analyzed how frequently constraints were violated. Figure 1 illustrates the learning curves for both groups. The probabilities of violating a constraint on the first and subsequent attempts were averaged over all students. The x-axis represents the attempt number (first, second and so on) when a student violated a constraint. The y-axis shows the probability of violating these constraints. The probability of making a mistake is initially higher for the experimental group than the control group even though not significantly. Figure 1 indicates that both groups learnt the constraints in a similar manner.

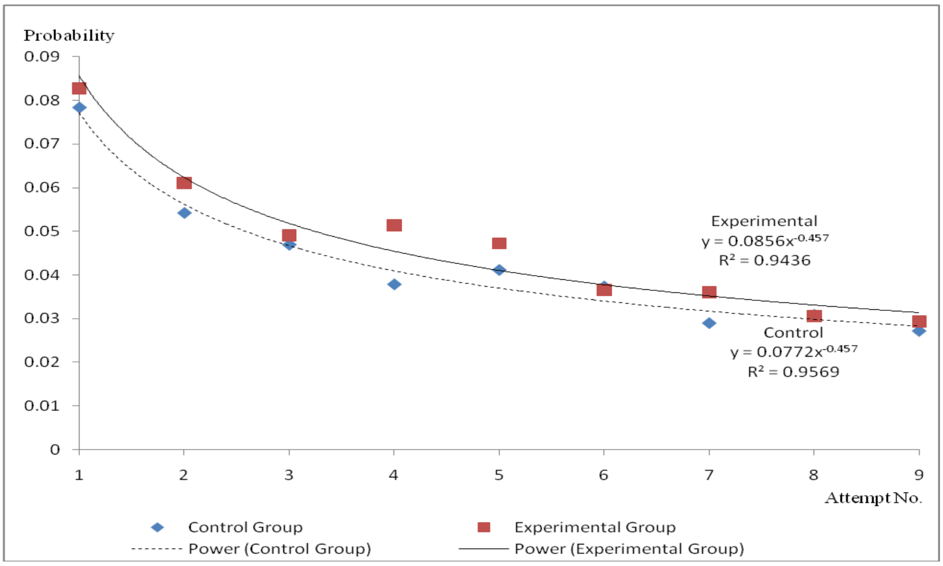

Fig 1: Probability of constraint violations – EER-Tutor study

We also investigated the number of constraints learnt by both groups. We used the first five attempts and the last attempts on each constraint to decide whether the status of the constraint changed from 'not known' to 'learnt' for a given student. If the probability of violating a constraint is below a pre-defined threshold then the constraint

[1] Normalised learning gain =learning gain/(1-pre-test score)

[2] Effect size = (Experimental Mean – Control Mean) /Standard Deviation of both groups

was deemed not known. Similarly, if the probability of violating a constraint is above the same pre-defined threshold then it was considered to be learnt. This analysis revealed that the experimental group learnt a significantly higher number of constraints than the control group (2.3 vs 1.2, p= 0.02).

Table 2 presents the subjective responses about various aspects of the dialogues. The impression about the quality of the dialogues and the ease of understanding the questions were similar between the groups. However there was clear evidence that the control group did not like having to go through the entire dialogue.

Table 2. Subjective responses about tutorial dialogues (sd given in parentheses)

Question	Likert scale	Control	Experimental	p
Quality of the dialogues	Poor to Excellent (1 to 5)	3.5 (1.0)	3. 7(0.8)	ns
Length of the dialogues	Too long to Too short (1 to 5)	2.6 (0.9)	3.2 (0.5)	0.002
Ease of understanding the questions	Very Hard to Very Easy (1 to 5)	3. (1.0)	3.4 (0.8)	ns

3. NORMIT Study

We conducted a study with NORMIT in September 2010 at the Victoria University of Wellington, which involved 20 volunteers from a database system engineering course in a single, 1-hour session. The objective and the experimental setup for this study are similar to that of EER-Tutor study. Pre-and the post-tests were designed to explore the system's effect on both the students' problem-solving abilities and explanation skills. Both pre- and post-tests had 4 questions each, of similar difficulty. Two questions requested students to solve very simple problems, and explain their solutions. The other two questions requested students to specify definitions of concepts. Some students have not completed the post-test. Table 3 reports some statistics about the 18 participants who completed both tests. Each group had 9 students.

Table 3: Some statistics from the NORMIT study (sd given in parentheses)

	Control (9)	Experimental (9)	p
Pre-test (%)	68.1 (30.0)	69.4 (29.4)	ns
Post-test (%)	72.2 (24.0)	86.1(15.9)	ns
Gain	4.2 (32.4)	16.7 (27.2)	ns
Interaction time (min)	60.1(24.7)	47.7 (16.8)	ns
Attempted Problems	7.1 (3.0)	5.9 (2.1)	ns
Solved problems	6.1 (3.0)	5.4 (2.0)	ns
Total Dialogues received	27.8 (14.6)	23.6 (11.3)	ns
Questions answered	55.7 (37.4)	23.9 (11.5)	0.01
% of correct answers	6.9 (4.1)	8.2 (4.7)	ns

There were no significant differences between the pre-test and post-test performances of the two groups, as well as between the gains. The performance of the experimental group increased significantly between pre- and post-test (paired t-test, t=1.84, p=0.052), while the improvement of the control group was not significant. The effect size for learning gains of the two groups is 0.4.

As the study was limited to a single lab session, the two groups spent a similar time interacting with the system. The groups attempted and solved a similar number of problems, and received a similar number of dialogues.

The control group participants answered significantly more questions than their peers, as was the case in the EER-Tutor study. This can be expected as the control group had to go through the entire dialogue every time a dialogue is given to the student. However, percentages of correct answers are similar for both groups.

Figure 2 presents the learning curves for both groups. The probability of making a mistake is initially higher for the experimental group than the control group even though not significantly. The learning curves indicate that the learning rate of the experimental group is higher than that of the control group. Similar to the EER-Tutor study, we also investigated the number of constraints learnt by both groups. There was no significant difference between the numbers of constraints learnt.

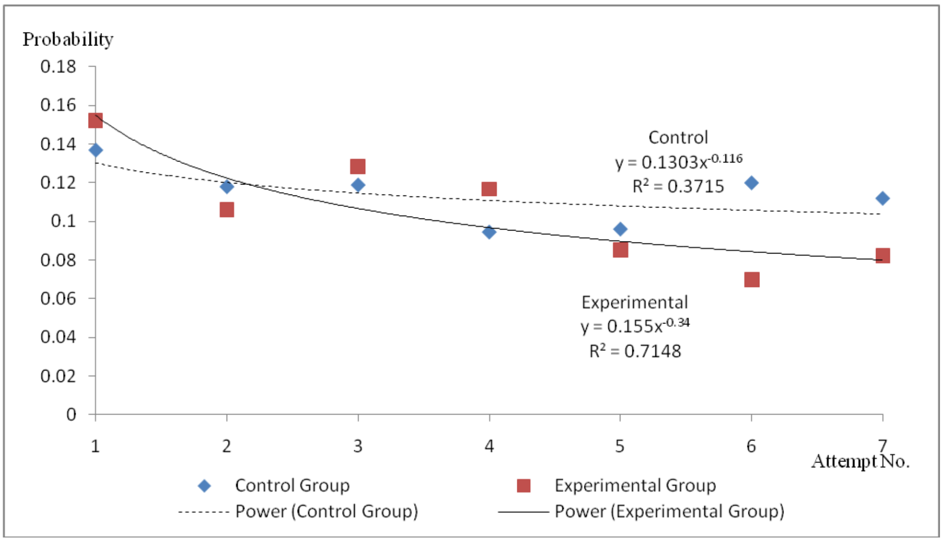

Fig 2: Probability of constraint violations – NORMIT study

We also explored the users' impressions about various aspects of tutorial dialogues using questionnaires (Table 4). The questions used for the EER-Tutor study were used here. The impression about the quality of the dialogues and the ease of understanding the questions were similar between the groups. Unlike the EER-Tutor study, there was no evidence from the control group that the non-adaptive dialogues were too long.

Table 4. Subjective responses about tutorial dialogues (sd given in parentheses)

Question	Likert scale	Control	Experimental	p
Quality of the dialogues	Poor to Excellent (1 to 5)	3.3 (0.5)	3.1(1.0)	ns
Length of the dialogues	Too long to Too short (1 to 5)	3.1 (0.8)	3.3(0.5)	ns
Ease of understanding the questions	Very Hard to Very Easy (1 to 5)	3.4(0.7)	3.1(0.7)	ns

4. Discussion and Conclusions

We presented how we evaluated the effectiveness of our model for supporting tutorial dialogues in two very different tasks. Our model facilitates adaptive dialogues based on a student's knowledge and their interaction with the dialogues. The dialogues discuss a student's mistake in the current context and the relevant domain concepts.

In EER-Tutor study the learning gain of the experimental group (that received adaptive dialogues) is significantly higher than the gain of their peers, with the effect size of 0.69. The experimental group also learnt a significantly higher number of constraints. These results strongly suggest that adaptive dialogues had a positive effect on learning database design. This is a significant result because (i) the difference between the two groups was minimal (i.e. the only difference was the adaptivity of the dialogues) and (ii) the study was limited to a single 2- hour session.

In the NORMIT study, there were no significant differences between the pre-test and post-test performances of the two groups, as well as between the gains. This might be due to the small number of participants (20 vs 65 in EER-Tutor study). However, we can observe similar trends in learning in both studies: significantly higher number of constraints learnt in EER-Tutor study, and a higher learning rate in NORMIT study by the respective experimental groups compared to their peers.

In both studies we used dialogues to discuss the errors in the problem-solving process, and not as the main activity to learn the domain knowledge. The task facilitated in EER-Tutor requires world knowledge about different real-world scenarios such as enrolling students in a university, or customers interacting with a bank. In the EER-Tutor study, the model was used to support dialogues in an ill-defined task with the well-defined domain theory. In the NORMIT study, dialogues facilitated learning a well-defined task with the well-defined domain theory. Therefore, our model has shown evidence of enhancing learning of a domain in the WDIT quadrant (well-defined domain, ill-defined task) and WDWT quadrant (well-defined domain, well-defined task) [6]. As the next step, we plan to explore the possibility of developing the model for a task such as essay writing or legal argumentation in the IDIT quadrant (Ill-defined domain, Ill-defined task).

The three highest levels of the error-hierarchy (the first component of the model) are domain-independent. The top level node is *All Errors*, which is then further divided into *Basic Syntax Errors* and *Errors dealing with the main problem solving activity*. The latter is further divided into (i) *Using an incorrect solution component type*, (ii) *Extra solution components,* (iii) *Missing solution components*, (iv) *Associations* and (v) *Failure to complete related changes*. Further divisions of these nodes and the node *Basic Syntax Errors* deal with domain-specific concepts. Even though tutorial dialogues consist of domain-specific prompts, the structure is domain-independent. Adaptation rules (the last component) which customise dialogue prompts are domain-independent except for the time period of inactivity the tutor waits before intervening.

We also investigated whether our model can be used in other domains. We tried to fit the errors from two different domains: logical database design and fraction addition into our model. Logical database design involves mapping high-level, conceptual ER schemas to relational schemas using the 7-step mapping algorithm [9]. We used the constraint-base of ERM-Tutor [10], a constraint-based tutor for teaching logical database design and developed the error hierarchy categorizing all the constraints. Then we explored whether we could develop dialogues for each type of error. All these were done on paper and the model could be developed for logical database design. We repeated the steps of (i) developing the error hierarchy using the constraints developed for fraction addition and (ii) developing dialogues for each type of error. The outcome of our attempt is a model that could be implemented to support dialogues in fraction addition. Therefore we have developed models for four different domains: (i) database design (ii) data normalization (iii) logical database design and (iv) fraction addition. The first two were implemented and evaluations indicate that the model can enhance learning the domain knowledge. The last two were done on paper and our attempt provides evidence that the model can be used in different domains.

For a newly created constraint-based tutor, developing our model to support dialogues involves (i) developing the error hierarchy to categorize the errors in the domain using the constraint-base (ii) designing the dialogues for each type of error and (iii) customizing the domain-dependent features (i.e. inactive time period) in the adaptation rules. Furthermore, even though this model was developed for constraint-based tutors, it can be used in any ITS with a problem-solving environment. In such ITSs, a student solution is evaluated and feedback is provided on errors regardless of the mechanism/methodology used for diagnosis. Therefore, the error hierarchy (the first component of the model) could be developed using the error types of that domain. Tutorial dialogues (the second component of the model) need to be written for each type of error based on the dialogue structure. The third component of the model, rules for adapting dialogues, are domain independent (except for the inactive time period), and can be used across domains.

The future work includes conducting a larger NORMIT study and exploring the possibility of developing a model for an ill-defined task in an ill-defined domain.

References

1. VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. :When are tutorial dialogues more effective than reading? Cognitive Science 31(1), 3-52, (2007).
2. Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., et al.: AutoTutor: A tutor with dialogue in natural language. Behavioral Research Methods, Instruments and Computers, *36*, 180–193,(2004).
3. Evens, M. and Michael, J., One-on-One Tutoring By Humans and Computers. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2006.
4. Aleven, V., Ogan, A. Popescu, O. Torrey, C., Koedinger, K.: Evaluating the Effectiveness of a Tutorial Dialogue System for Self-Explanation. In: Lester, J. et al. (Eds.) ITS2004, LNCS, vol. 3220, pp 443-454, Springer-Verlag, Berlin (2004).
5. Weerasinghe, A., Mitrovic, A.: Facilitating Deep Learning through Self-Explanation in an Open-ended Domain. Knowledge-based and Intelligent Tutoring Systems 10(1), 3-19 (2006).
6. Mitrovic, A., Weerasinghe, A.: Revisiting the Ill-Definedness and Consequences for ITSs. Dimitrova, V. et al. (Eds.) Proc. Artificial Intelligence in Education, Frontiers in Artificial Intelligence and Applications, vol. 200, pp. 375-382 (2009).
7. Weerasinghe, A., Mitrovic, A., Martin, B.: Towards Individualized Dialogue Support for Ill-Defined Domains IJAIED, Special Issue on Ill-Defined Domains, 19(4): pp. 357-379 (2009).
8. Mitrovic, A., Martin, B., Suraweera, P: Intelligent Tutors for All: Constraint-based Modeling Methodology, Systems and Authoring. IEEE Intelligent Systems 22(4), 38-45 (2007).
9. Elmasri, R., Navathe, S., Fundamentals of Database Systems (5th ed.). Boston: Addison-Wesley (2007).
10. Milik, N., Marshall, M., Mitrovic, A.: Teaching logical database design in ERM-Tutor.: Ikeda, M., Ashley, K. (Eds.) Proc. of ITS2006, pp. 707-709 (2006).