S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Evaluating the Effectiveness of Adaptive Tutorial Dialogues in EER-Tutor

Amali WEERASINGHE, Antonija MITROVIC, Martin VAN ZIJL, Brent MARTIN
Intelligent Computer Tutoring Group, University of Canterbury, New Zealand
amali.weerasinghe@pg.canterbury.ac.nz

Abstract: Researchers have long been interested in tutorial dialogues as they are considered to be one of the critical factors contributing to the effectiveness of human one-on-one tutoring. We discuss an evaluation study that investigates the effectiveness of adaptive tutorial dialogues in database design. EER-Tutor, a database design tutor was enhanced to facilitate adaptive tutorial dialogues. 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 indicated that the experimental group participants learned significantly more than their peers. The subjective responses indicated no difference in their impression towards the quality of the dialogues and the understandability of the questions. However there was clear evidence that the control group did not like having to go through the entire dialogue before resuming problem-solving.

Keywords: tutorial dialogues, constraint-based tutors, adaptive dialogues

Introduction

One-on-one human tutoring is wildly considered to be the most effective form of instruction [2]. Students' learning gains have been increased by two standard deviations when tutored by human tutors compared to traditional classroom instruction. This has inspired researchers to explore how the effectiveness of human tutoring strategies can be incorporated into intelligent tutoring systems. One of the critical factors contributing to the effectiveness of human tutoring is the conversational aspect of the instruction. Dialogues provide opportunities for students to reflect on their existing knowledge and to construct new knowledge. Some of the dialogue-based tutoring systems that have been developed are Why2-Atlas [3], Auto Tutor [3], CIRCSIM-Tutor [4], Geometry Explanation Tutor [1] and KERMIT-SE [9]. Why2-Atlas and AutoTutor use dialogues as the main activity to help students learn the domain knowledge. The other systems provide problem-solving environments as the main activity and use tutorial dialogues as a way of remediating errors in the student solutions. For example, CIRCSIM-Tutor is a natural language (NL) 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 instructional 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.

Our long-term goal is to develop a general model for supporting dialogues across domains. Since we previously implemented dialogues for EER-Tutor [5], the initial work on this project started with the same system. Based on the findings of two Wizard-of-Oz studies [7, 8], we developed a model to support dialogues. Our model consists of three parts: an error hierarchy, tutorial dialogues and rules for adapting them. The error hierarchy categorizes all the error types in a domain. At the lowest level an error type is associated with one or more violated constraints, which form leaves of the hierarchy. 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 the correct option from a list provided by the tutor. For a detailed discussion of the model see [7].

The next section presents the details of the evaluation study. Section 2 presents the results followed by conclusions.

1. The 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. The participants were randomly assigned to two groups (experimental and control). The experimental group received adaptive support based on our model. The control group was given non-adaptive support in which two different students with different knowledge levels received the same dialogue. Differences between the two groups were: (i) Dialogue selection (ii) Dialogue prompts and (iii) Additional support.

Dialogue selection: The dialogue selection for the control group was based on a 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 were ordered from simpler to more complicated errors, the depth-first search results in the simplest error to initiate a dialogue. For instance, Figure 1 presents the dialogue that a control group participant receives for the submitted solution. It contains multiple errors: (i) ROOM should be represented as a weak entity instead of a regular entity (ii) Attributes are missing from the entities HOTEL, EMPLOYEE and ROOM (iii) Cardinality between HOTEL and WORKS_FOR is incorrect etc. The error selected for discussion was that ROOM was modelled as a regular entity. Now consider an experimental group participant with an identical student model to the previous student submitting the same solution. Figure 2 represents the dialogue to be received. This dialogue focuses on the incorrect cardinality between EMPLOYEE and HOTEL. This is because cardinality was identified as the most difficult concept based on his/her student model. (i.e. the error this student will most likely to make in the next attempt).

Dialogue prompts: 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 a result, the same solution submitted by two different students with different knowledge levels in the control group received identical dialogues. For instance, the prompt received by the control group participant discusses the domain concept related to the error selected for discussion (EERTutor1 in Figure 3(a)). We call this type of prompt a problem-independent prompt as it focuses on the relevant domain concept [7]. The entire dialogue is given in Figure 3(a). In contrast, the prompt received by the experimental group participant discusses the selected error in the context of the current problem (EERTutor3 in Figure 3(b)). This type of prompt is called problem-dependent prompt [7]. This error was chosen for discussion because his/her student model identifies that cardinality is the most difficult concept at this moment. He/She receives a problem-dependent prompt (instead of a

problem-independent prompt) because this is the first time this mistake is made during the current session. If he makes this type of error repeatedly, he will be given the problem-independent prompt (EER-Tutor1 in Figure 3(b)).

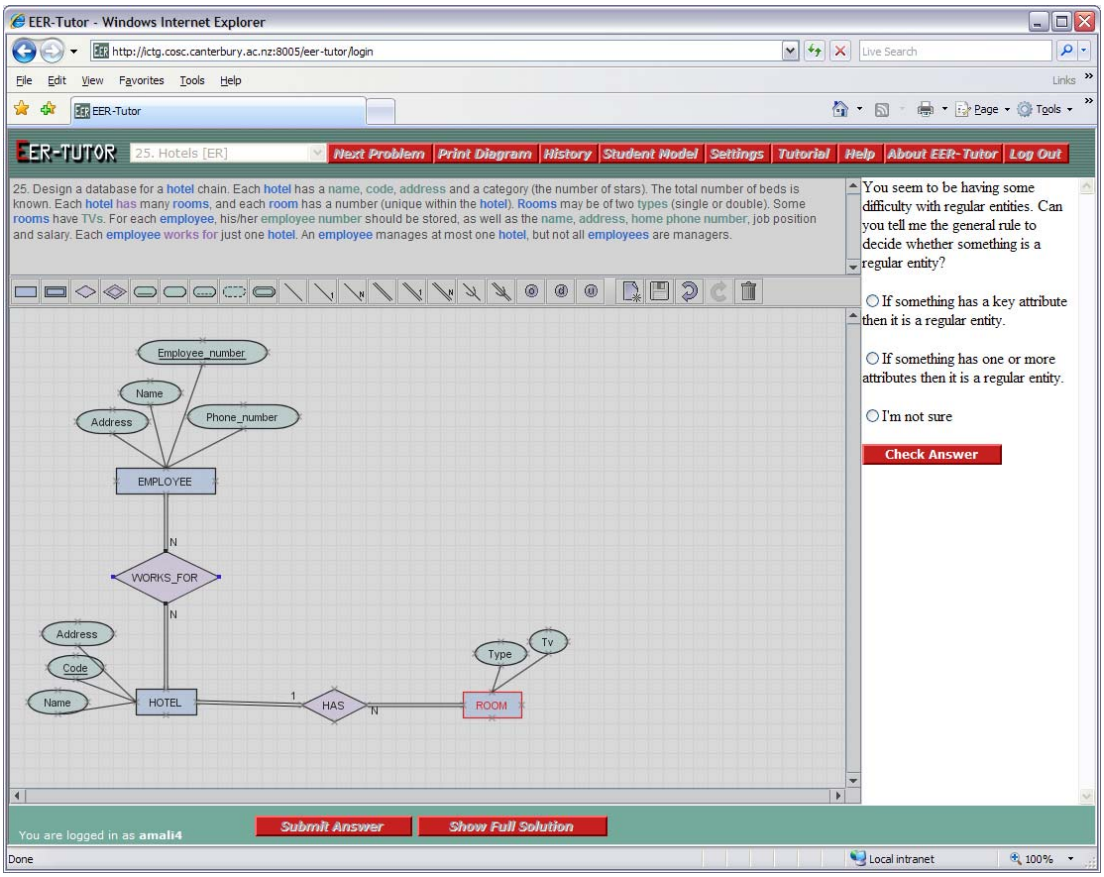

Fig. 1. The dialogue received by a control group participant

Additional support: When an experimental group participant abandons a problem (i.e. changes a problem without submitting at least once) 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: (i) pre-test (ii) interactions with EER-Tutor (iii) post-test (iv) questionnaire.

Pre- and post-tests: Pre-tests were used to determine the participants' knowledge before interacting with the system and also to determine whether the knowledge between the experimental and control was significantly different. Both pre- and post-tests had 6 questions each. The questions in the pre- and post-tests were 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. Three other questions were aimed to understand the effect

the system had on students' explanation skills. The remaining two questions asked about declarative knowledge.

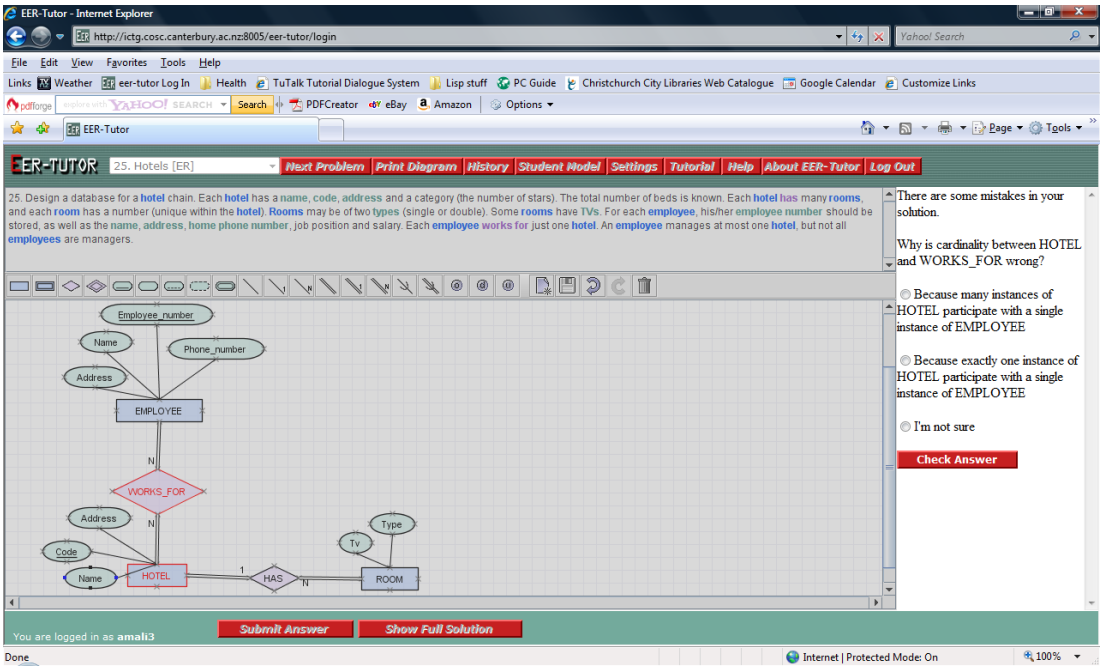

Fig 2. The dialogue received by an experimental group participant

Questionnaire: There were eight questions aimed at understating participants' view of different aspects of dialogues. Five questions had Likert scales (ranging from 1 to 5) discussing the quality, the length and the prompts in the dialogues. Participants were also given an opportunity to explain how the dialogues helped them in their learning. Some questions focused on suggestions to improve the dialogues and EER-Tutor in general.

Procedure: The participants used EER-Tutor for the first time in their regular lab sessions during the third week of the course, by which time they had been introduced to ER modeling.
They were randomly divided into the control and experimental groups. The duration of the study was limited to a single 2-hour session. At the beginning of the session students were given about 10 minutes to complete the pre-test. Then they started solving ER modelling problems using EER-Tutor. Towards the end of the session, they were given 10 minutes to complete the post-test and 5 minutes to answer the questionnaire.

2. Results and Discussion

Out of 104 students enrolled in the course, 77 participated in the study. Some students did not complete the post-test. Table 1 reports some statistics about the 65 participants who completed both pre-and post-tests. There were 31 participants in the experimental group and 34 in the control group. There was no significant difference on the pre-test performances of the two groups. The post-test performance of the experimental group was significantly better compared to their peers who received non-adaptive dialogues Both the learning gain[1]

[1] Learning gain = post-test score – pre-test score

and the normalised learning gain[2] of the group who received adaptive dialogues was also significantly higher than the gains of the control group.

EERTutor1: You seem to be having some difficulty with regular entities. Can you tell me the general rule to decide whether something is a regular entity? ***Student1***: *I don't know* **EERTutor2:** Let's try a simple thing. What is used to uniquely identify instances of a regular entity ***Student2***: key a*ttribute* **EERTutor3 :** Great ! Now tell me why modelling *ROOM* as a regular entity is incorrect? ***Student3***: *I'm not sure* **EERTutor4**: Let's see. Check carefully whether you can find a key attribute for *ROOM*. {When Continue button is pressed} Now please select the key attribute for *ROOM* ***Student4***: *No key attribute for ROOM* **EERTutor5**: Great! Now tell me the best way to model *ROOM*? ***Student5***: *I need more help* **EERTutor6:** Ok. As *ROOM* does not have a key attribute it cannot be a regular entity. So it should be a weak entity. **EERTutor7**: Great ! try to answer this question so that you'll remember what you learnt. …….. uniquely identifies instances of a …… ***Student6***: key *attribute, regular entity* **EERTutor8**: Excellent, go ahead and do the changes. Submit the solution when you're done.	**EERTutor1**: Looks like you have trouble with cardinality. Let's focus on cardinality now. What do you represent by cardinality on the E1 side between entities E1 and E2 in a binary relationship? ***Student1***: *I don't know* **EERTutor2:** Here's a simple question to try. How do we model connections between entities? ***Student2***: *Using relationships* **EERTutor3 :** Great ! Why is the cardinality between *HOTEL and WORKS_FOR* is wrong? ***Student3***: *I'm not sure* **EERTutor4**: Let's see. Try to figure out how many instances of *HOTEL* participates with a single instance of *EMPLOYEE*? {When Continue button is pressed}Now tell me how many instances of *HOTEL* participates with a single instance of *EMPLOYEE*? ***Student4***: Exactly one **EERTutor5:** Great! Guess you know how to make change now. Before that, try to answer this question. What is the correct question to ask when deciding the cardinality on the E1 side between entities E1 and E2 in a binary relationship? ***Student5***: How many instances of E1 participates with a single instance of E2? **EERTutor6**: Excellent job, go ahead and do the changes. Submit the solution when you're done.

Fig. 3 A hypothetical dialogue for EER-Tutor received by a (a) control group participant (b) experimental group participant

As the study was limited to a single lab session, the two groups spent a similar time interacting with the system. There were also no significant differences between the number of attempted and solved problems. The total number of dialogues, the total number of single-level dialogues (some dialogues are limited to a single feedback message as they discuss simple errors) and the total number of multi-level dialogues were also similar for the two groups.

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, percentage number of correct answers was similar for both groups. There were no significant differences on the total number of questions answered incorrectly or the questions with a *More Help* request (i.e one of the options available was *I don't know* or *I need more help* which resulted in presenting the relevant information to the

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

student). Also there was no significant difference on the percentage of questions that requested more help. However, it is interesting to note that the experimental group has provided a significantly higher percentage of incorrect answers. Further analysis is required to understand the cause for this.

Table 1. Some statistics from the 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
No. of constraints learnt	1.2 (1.5)	2.3 (2.3)	0.02
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
Single-level dialogues seen	2.1(3.0)	1.9 (2.7)	ns
Multi-level dialogues seen	10 (6.8)	12.1(7.2)	ns
Total number of questions answered	34.4 (25)	23.6 (14.6)	0.01
Total number of questions answered correctly	23.3 (17.9)	14 (10.4)	0.006
% number of questions answered correctly	61.4(23.1)	59(16.9)	ns
Total number of questions answered incorrectly	9.1 (8.3)	7.3 (4.3)	ns
% number of questions answered incorrectly	23.7(12.9)	31.8(15)	0.01
Total number of questions with a *More Help* request	2.1 (3.5)	2.4 (3.5)	ns
% number of questions answered with a *More Help* request	6.1(6.9)	9.22(11.4)	ns

Effect size[3] is a standard way to compare the results of one pedagogical experiment to another. It indicates how much more the experimental group has learnt compared to the control group? The effect size (Cohen's d) for learning gains of the two groups is 0.69 (the effect size based on the normalized gain is 0.51). This is comparable to the study with SQL-Tutor conducted in a similar setting in a single 2-hour session [6]. An effect size of 0.66 was reported for that study for the students who used SQL-Tutor compared with those who did not use the tutor. The effect size obtained here is therefore remarkable because the only difference between the two groups was the adaptivity of the dialogues.

2.1 Learning Curves

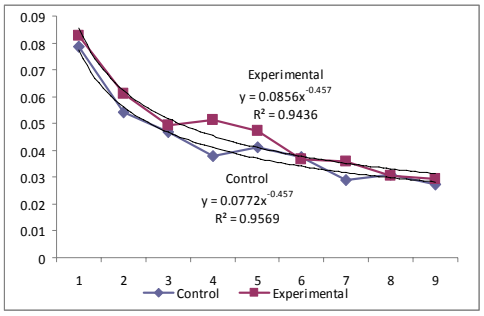

Fig 4: Probability of violating a constraint

In order to investigate how the students in both groups learnt the database design concepts in terms of constraints we analyzed how frequently constraints were violated. Figure 4 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

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

though not significant. Figure 4 indicates that both groups learnt the constraints in a similar manner. Both learning curves have a good fit to the power curve, indicating that the transferability of learning is high for both groups

We also investigated the number of constraints learnt by both groups. For each constraint in a student model, the first 5 attempts and the last 5 attempts during which a constraint was relevant was considered. If the probability of violating a constraint was reduced by 0.7 during the last 5 attempts, then that constraint 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).

2.2 Subjective Responses

Table 2 presents the subjective responses about various aspects of the dialogues. The starting and the ending points of the Likert scale had descriptive labels and the middle points had only numeric labels. For instance, when asking about overall quality of the dialogues, the starting and the ending labels were Poor (1) and Excellent (5) The points 2, 3 and 4 were indicated on the scale. 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 (standard deviation reported in parentheses)

Question	Likert scale	Control	Experimental	p
Overall 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(1.0)	3.4(0.8)	ns

3. Conclusions

We discuss an evaluation study that investigates the effectiveness of adaptive tutorial dialogues in EER-Tutor. 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 model. The study was conducted in their regular lab sessions and was limited to a single 2-hour session. At the end of the session the performance of the experimental group participants increased significantly more than their peers with an effect size of 0.69. The experimental group also learnt a significantly higher number of constraints than the control group. These results strongly suggest that the adaptive dialogues had a positive effect on learning database design. These results are significant because (i) the difference between the two groups was minimal (i.e. the only difference was the adaptivity of the dialogues) and (ii) the duration of the study was limited to a single 2- hour session. The subjective responses indicated no difference in their impression towards the quality of the dialogues and/or the understandability of the questions. However there was clear evidence that the control group did not like having to go through the entire dialogue before resuming problem-solving.

The participants were given the opportunity to interact with the system after this study. These interactions will be analysed to see how motivated they were to use EER-Tutor in their own time. Also we plan to use performance on their assignment which requires them to design a complex data model as a delayed post-test to investigate their improvement in their knowledge in database design.

Acknowledgements:

We would like to thank Benedict du Boulay from University of Sussex for his help with the evaluation study.

References

[1] Aleven, V., Ogan, A. Popescu, O. Torrey, C., & Koedinger, K. R. (2004). Evaluating the effectiveness of a Tutorial Dialogue System for Self-Explanation, Lester, J. C., Vicario, R.M., Papaguacu, F. (eds.) *Proceedings of ITS2004*, (pp 443-454). Alagoas, Brazil: Springer-Verlag.

[2] Bloom, B. (1984) The 2-sigma problem: The search of group instruction as effective as one-to-one tutoring, *Educational Researcher* 13, 3-16.

[3] Grasser, A.C., VanLehn, K., Rose, C.P., Jordan P. W., & Harter, D. (2001) Intelligent Tutoring Systems with Conversational Dialogue *AI magazine*, 22(4), 39-51.

[4] Millis, B., Evens, M., & Freedman, R. (2004). Implementing Directed Lines of Reasoning in an Intelligent Tutoring System using the Atlas Planning Environment, *Proceedings of the International Conference on Information Technology*: *ITCC 2004*, (pp. 729-733). Nevada-USA: IEEE Computer Society.

[5] Mitrovic, A., Martin, B. & Suraweera, P. (2007). Intelligent Tutors for all: Constraint-based modeling methodology, systems and authoring, *IEEE Intelligent Systems*, 22(4), 38-45.

[6] Mitrovic, A., Martin, B., & Mayo, M. (2002) Using Evaluation to Shape ITS Design: Results and Experiences with SQL-Tutor. *User Modeling and User-Adapted Interaction*, 12 (2-3), 243-279.

[7] Weerasinghe, A., & Mitrovic, A. (2008). A Preliminary Study of a General Model for Supporting Tutorial Dialogues. *Proceedings of the International Conference in Computers in Education*, (pp. 125-132). Taipei, Taiwan: Asia-Pacific Society for Computers in Education.

[8] Weerasinghe, A., & Mitrovic, A. (2006). Individualizing Self-Explanation Support for Ill-Defined Tasks in Constraint-based tutors, Aleven V. Ashley, K., Lynch, C. and Pinkwart, N. (eds.), Workshop on Intelligent Tutoring Systems for ill-defined domains at ITS2006, (pp. 55-64). Jhongli, Taiwan.

[9] Weerasinghe, A., & Mitrovic, A. (2006). Facilitating Deep Learning through Self-Explanation in an Open-ended Domain, *Knowledge-based and Intelligent Tutoring Systems* 10(1), 3-19.