Adapting to Student Uncertainty Improves Tutoring Dialogues

Kate FORBES-RILEY [1], Diane LITMAN
University of Pittsburgh, Pittsburgh, PA, 15260, USA

Abstract. This study shows that affect-adaptive computer tutoring can significantly improve performance on learning efficiency and user satisfaction. We compare two different student uncertainty adaptations which were designed, implemented and evaluated in a controlled experiment using four versions of a wizarded spoken dialogue tutoring system: two adaptive systems used in two experimental conditions (*basic* and *empirical*), and two non-adaptive systems used in two control conditions (*normal* and *random*). In prior work we compared learning gains across the four systems; here we compare two other important performance metrics: learning efficiency and user satisfaction. We show that the *basic* adaptive system outperforms the *normal* (non-adaptive) and *empirical* (adaptive) systems in terms of learning efficiency. We also show that the *empirical* (adaptive) and *random* (non-adaptive) systems outperform the *basic* adaptive system in terms of user perception of tutor response quality. However, only the *basic* adaptive system shows a positive correlation between learning and user perception of decreased uncertainty.

Keywords. computer tutor, affect adaptation, spoken dialogue, empirical evaluation

1. Introduction

Student uncertainty is of increasing interest in the computer tutoring community, due both to its theorized relationship to correctness and learning, and to the fact that it is among the most frequently occurring student affective states during computer tutoring (e.g. [1,2,3,4,5]).[2] Promising results have been reported on correlating uncertainty and learning [1,5] and on manually annotating (e.g. [5,8]) and automatically detecting (e.g. [8,9,3]) uncertainty in computer tutoring dialogues. However, little research has yet focused on developing, implementing, and evaluating contentful computer tutor responses to student uncertainty.

In this paper we show that responding to student uncertainty can significantly improve performance in a wizarded spoken dialogue tutoring system. We analyze the impact on performance of two student uncertainty adaptations which were designed, implemented and evaluated in a controlled experiment using four versions of our system: two adaptive systems used in two experimental conditions (*basic* and *empirical*), and two non-adaptive systems used in two control conditions (*normal* and *random*). Else-

[1] Corresponding Author: forbesk@cs.pitt.edu

[2] Although uncertainty is not among the "big 6" set of "basic" emotions (anger, disgust, fear, happiness, sadness, surprise) [6], computer tutoring researchers have found this set needs to be supplemented or even replaced to describe the range of emotions and attitudes actually displayed by users (e.g. [7]). We thus use the term "affect" for emotions and attitudes displayed by students using our spoken dialogue tutoring system.

where we compare learning gains across the four systems [10]. Here we investigate the differences between the four systems with respect to two other important tutoring system performance metrics: learning efficiency (measured as learning gain divided by time on task and total student turns) and user satisfaction (measured by user survey questions). We find that the *basic* adaptive system outperforms the *normal* system significantly, and the *empirical* adaptive system as a trend, in terms of learning efficiency. We also find that the *empirical* adaptive system and the *random* system both outperform the *basic* adaptive system in terms of user perception of tutor response quality. However, only the *basic* adaptive system shows a positive correlation between learning and user perception of decreased uncertainty.

2. ITSPOKE: Affect-Adaptive and Non-Adaptive Versions

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) is built on top of the Why2-Atlas text-based tutor [11]. ITSPOKE tutors 5 qualitative physics problems via spoken dialogue. The dialogues have a Tutor Question - Student Answer - Tutor Response format, implemented with a finite state dialogue manager, and consist of a series of questions about the topics needed to solve the problem. In our original (non-adaptive) version of ITSPOKE, tutor responses (states) depended only on the correctness of student answers (transitions between states). In our two affect-adaptive versions of ITSPOKE, tutor responses depend on both the correctness and the uncertainty of student answers.

2.1. Basic Affect-Adaptive ITSPOKE

In *basic* affect-adaptive ITSPOKE, tutor responses are determined as follows: If the student answers *correctly and without uncertainty*, ITSPOKE responds with Correctness feedback (e.g. “Right”) then moves on to the next question. If the student answers *incorrectly with or without uncertainty*, ITSPOKE responds in one of two forms (both begin with Incorrectness feedback, e.g. “Well...”): 1) For incorrect answers to easier questions, ITSPOKE “Bottoms Out”, i.e., provides the correct answer with a brief statement of reasoning. 2) For incorrect answers to harder questions, ITSPOKE initiates a “Remediation Subdialogue”, i.e., one or more questions that walk through a more complex line of reasoning.. Finally, if the student answers a question *correctly but with uncertainty*, ITSPOKE treats the answer as if it were incorrect, i.e. gives the same response it would give if the answer were incorrect (except the response begins with Correctness feedback).

Impasse State:	InonU	IU	CU	CnonU
Severity Ranking:	most	less	least	none

Figure 1. Different Learning Impasse State Severities

This *basic* uncertainty adaptation is derived from tutoring theory relating uncertainty to incorrectness and learning. VanLehn et al. [4] view both uncertainty and incorrectness as signals of “learning impasses”; i.e., opportunities for the student to learn. We further observed that to resolve an impasse, the student must first perceive that it exists. Incorrectness and uncertainty differ in this perception. Incorrectness simply signals the student has reached an impasse, while uncertainty signals the student perceives s/he has reached an impasse. Based on this, we associated each combination of binary uncertainty (**U**, **nonU**) and correctness (**I**, **C**) with an “impasse severity”, as in Figure 1. A student is not experiencing an impasse if s/he is correct and not uncertain about it (CnonU). A student is experiencing the most severe type of impasse if s/he is incorrect and not aware of it (InonU). IU and CU answers indicate impasses of lesser severity: the student is incorrect but aware s/he may be, and the student is correct but uncertain if s/he is, respectively. In Forbes-Riley et al. [12] we show empirical support for distinguishing impasse severities.

The hypothesis underlying our *basic* uncertainty adaptation is that performance will improve if ITSPOKE provides *the same additional content* to remediate *all* impasses (CU, IU, InonU). The original non-adaptive ITSPOKE provides additional content to resolve incorrectness impasses (IU and InonU), but it ignores one type of uncertainty impasse (CU). The *basic* affect-adaptive ITSPOKE provides the same response to CU, IU, and InonU answers. Figure 2 illustrates the *basic* uncertainty adaptation. As shown, the CU answer in **STU1** receives a Remediation Subdialogue (only the first question is shown) in **TUT2**, along with feedback acknowledging correctness. This response would also have been given if **STU1** were incorrect regardless of uncertainty.

TUT1: Now let's talk about the net force exerted on the truck. By the same reasoning that we used for the car, what's the overall net force on the truck equal to?
STU1: The force of the car hitting it?? *[CU]*
TUT2: Fine. *[FEEDBACK]* We can derive the net force on the truck by summing the individual forces on it, just like we did for the car. First, what horizontal force is exerted on the truck during the collision? *[REMEDIATION SUBDIALOGUE]*

Figure 2. Example of *Basic* Uncertainty Adaptation (for CU Answers)

2.2. Empirical Affect-Adaptive ITSPOKE

Our *empirical* uncertainty adaptation revises our *basic* adaptation based on empirical analyses of human tutor responses. The hypothesis underlying our *empirical* adaptation is that performance will improve if ITSPOKE provides the same additional content for all learning impasses (CU, IU, InonU), but varies *the dialogue act* used to present this content and varies *the (In)Correctness feedback* based on the answer's learning impasse severity (Figure 1). Examples are shown in Figure 3.

Our feedback variations were based on prior results in the computer tutor literature showing that human tutor-derived empathetic system responses can positively impact performance (see Section 4). Our new feedback phrases responded to the specific type of learning impasse signaled by the answer. In other words, the feedback acknowledged both the propositional content ((in)correctness) and the affective content (uncertainty).[3]

Our dialogue act variations were based on our own empirical analysis of human responses to uncertainty in a human tutoring corpus that corresponds to our ITSPOKE corpora. This analysis is discussed in [10]. Briefly, our human tutor responded with a question significantly less than expected after a CU answer. Since the original ITSPOKE response to an incorrect answer will either be a Remediation Subdialogue (beginning with a question) or a Bottom Out (declarative statement), we interpreted our human tutor result as follows: If the original response is a Bottom Out, use it after CU answers. Else author and use a new Bottom Out version of the existing Remediation Subdialogue.

Our human tutor also responded with a Bottom Out significantly more than expected after an IU answer and significantly less after an InonU answer. We interpreted this result as follows: 1) If the original response is a Bottom Out, use it after IU answers. Else

[3] We did not explicitly acknowledge the lack of an impasse in CnonU answers because a pilot study showed this feedback quickly became annoying (e.g. "That's correct and you don't sound uncertain so let's move on.")

author a new Bottom Out version of the existing Remediation Subdialogue, and use *both* responses. The new Bottom Out shows the final solution, then the existing Remediation Subdialogue walks the student through the steps to this solution. 2) If the original response is a Remediation Subdialogue, use it after InonU answers. Else author a new question-based version of the existing Bottom Out, and use *both* responses. The new (typically easy) question helps perceive the impasse, then the Bottom Out reinforces and explains the correct answer.

STU1: The car's force hitting the truck?? *[CU]*
TUT2: That's exactly right, but you seem unsure, so let's sum up. *[FEEDBACK]* The net force on the truck is equal to the impact force on it. We can prove this just like we did for the car. First, we know gravity and the normal force on the truck must cancel each other, otherwise the truck would not be at rest vertically. Second we know that the impact force is the only horizontal force exerted on the truck. *[NEW BOTTOM OUT]*

STU1: I don't know for sure. *[IU]*
TUT2: That's not correct, but don't worry. You seem to be aware of your mistake. Let's resolve it. *[FEEDBACK]* The net force on the truck is equal to the impact force on it. Let's walk through this answer step by step. *[NEW BOTTOM OUT]* We can derive the net force on the truck by summing the individual forces on it, just like we did for the car. First, what horizontal force is exerted on the truck during the collision? *[REMEDIATION SUBDIALOGUE]*

STU1: Zero. *[InonU]*
TUT2: I'm sorry, but I see an error in your answer. Let's fix it. *[FEEDBACK]* We can derive the net force on the truck by summing the individual forces on it, just like we did for the car. First, what horizontal force is exerted on the truck during the collision? *[REMEDIATION SUBDIALOGUE]*

Figure 3. *Empirical* Uncertainty Adaptation for **CU, IU, and InonU** Answers to **TUT1** in Figure 2

2.3. Controlled Experiment

We conducted a controlled experiment investigating the effectiveness of our uncertainty adaptations in ITSPOKE using a Wizard of Oz scenario (WOZ), in which a few ITSPOKE components were replaced by a human "Wizard": The Wizard performed speech recognition, language understanding, and uncertainty annotation, for each student answer. In this way, we tested the upper bound performance of our adaptations without any potentially negative impact of automated versions of these tasks.

The experiment had two control and two experimental conditions, each with 20-21 subjects. The *normal* and *random* control conditions both used the original non-adaptive ITSPOKE; however the *random* condition treated a percentage of randomly selected correct answers as incorrect, to control for the additional tutoring in the experimental conditions. The *basic* and *empirical* experimental conditions used the *basic* and *empirical* adaptive ITSPOKEs, respectively. Subjects were native English speakers who had never taken college physics, and were randomly assigned to the 4 conditions, except conditions were gender- and pretest-balanced. Each subject: i. read a background physics text; ii. took a pretest; iii. worked 5 problems with an ITSPOKE version (each problem yields 1 dialogue); iv. took a survey (Figure 4); v. took a posttest.

3. Evaluating Affect-Adaptive and Non-Adaptive ITSPOKE

In Forbes-Riley and Litman [10] we show that while students learned significantly over all conditions, the amount learned depends on condition. Here we evaluate how the affect-adaptive and non-adaptive conditions differed with respect to two additional types of evaluation metric: *learning efficiency* and *user satisfaction*. Learning efficiency refers to the amount of learning achieved in a given amount of tutoring (e.g., as in [13]). Because students take different amounts of time per turn, we measured learning efficiency for each student in two ways: 1) NLG/total tutoring time (NLG = normalized learning gain = (posttest-pretest)/(1-pretest)); 2) NLG/total student turns over all 5 dialogues.

SQ1: It was easy to learn from the tutor.
SQ2: The tutor didn't interfere with my understanding of the content.
SQ3: The tutor believed I was knowledgeable.
SQ4: The tutor was useful.
SQ5: The tutor was effective on conveying ideas.
SQ6: The tutor was precise in providing advice.
SQ7: The tutor helped me to concentrate.
SQ8: The tutor responded effectively after I was incorrect about the answer to a question.
SQ9: The tutor responded effectively after I was correct about the answer to a question.
SQ10: The tutor responded effectively after I was uncertain about the answer to a question.
SQ11: The tutor responded effectively after I was certain about the answer to a question.
SQ12: The tutor's responses decreased my uncertainty about my understanding of the content.
SQ13: It was easy to understand the tutor speech.
SQ14: I knew what I could say or do at each point in the conversations with the tutor.
SQ15: The tutor worked the way I expected it to.
SQ16: Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.
ALMOST ALWAYS (5), OFTEN (4), SOMETIMES (3), RARELY (2), ALMOST NEVER (1)

Figure 4. ITSPOKE Survey

User satisfaction refers to students' subjective perceptions of likability, ease of use, effectiveness, etc, which we measured with the survey in Figure 4. Students rated their degree of agreement with each statement on a scale of 1 to 5. Questions 1-7 and 13-16 were used in our prior ITSPOKE studies (see [10]); questions 8-12 were added for this study to address effectiveness relating to correctness and uncertainty. For user satisfaction metrics, we used total survey score as well as the rating for each question.

For each evaluation metric, we ran a one-way ANOVA with condition as the between-subject factor, along with a planned comparison for each pair of conditions, hypothesizing the following performance ranking: *empirical* > *basic* > *random* > *normal*.

The ANOVAs revealed significant differences between conditions in both measures of learning efficiency: NLG/time (F(3,77) = 3.56, p=0.02) and NLG/turns: (F(3,77) = 3.09, p=0.03) and for one user satisfaction survey question: SQ13 (F(3,77) = 2.69, p=0.05). Table 1 shows the significant results ($p \leq 0.05$) of the planned comparisons for these three metrics. The first column shows the metric, and the remaining columns list the condition, its mean and standard deviation, the condition with which a difference is found, and the direction (> or <) and significance of this difference.

As shown, *basic* has significantly higher learning efficiency than both *normal* and *empirical*. These results suggest that given the same amount of tutoring time, *basic* affect-adaptive ITSPOKE will effect more student learning than either our original non-adaptive ITSPOKE or *empirical* affect-adaptive ITSPOKE.[4]

As shown, both *empirical* and *random* significantly outperform *basic* for SQ13. This suggests that students perceive the tutor speech in *basic* affect-adaptive ITSPOKE as hard to understand. This may reflect students' confusion during the dialogues as to why *basic* affect-adaptive ITSPOKE was treating their correct+uncertain answers as incorrect. Because they were always already uncertain at this point, this treatment may have

[4] Our learning efficiency results also hold when computed using RLG (RLG = raw learning gain = posttest-pretest); i.e., the same results hold for RLG/time and RLG/turns.

confused them. In contrast, *empirical* affect-adaptive ITSPOKE explained that it would provide further discussion due to their uncertainty.

Table 1. Planned Comparison Results Showing Differences in Metrics Across Condition

Metric	Condition	Mean	SD	Diff	p
Learning Efficiency: NLG/total time (min)	*normal*	0.0100	0.0064	$<$ *basic*	0.004
	random	0.0135	0.0076	-	
	basic	0.0161	0.0057	-	
	empirical	0.0107	0.0069	$<$ *basic*	0.013
Learning Efficiency: NLG/total student turns	*normal*	0.0050	0.0028	$<$ *basic*	0.010
	random	0.0067	0.0039	-	
	basic	0.0075	0.0023	-	
	empirical	0.0053	0.0030	$<$ *basic*	0.023
SQ13	*normal*	3.90	0.77	-	
	random	4.15	0.75	$>$ *basic*	0.016
	basic	3.50	1.00	-	
	empirical	4.15	0.81	$>$ *basic*	0.016

Although our results suggest that our students don't express a strong preference for our affective systems, we hypothesized there might be a relationship between preference and learning; for example, subjects who prefer an affective system might learn more from it than those who do not. To investigate this relationship, we ran a Pearson's correlation between each user satisfaction metric and posttest (controlled for pretest) over all students and within each condition. This relationship has also been investigated in prior computer tutor research, with mixed results. For example, Moreno et al. [14] find significant differences in student preferences for various animated computer tutors, but find no relationship between these preferences and learning over all students. Rotaru [15] finds correlations between user satisfaction and learning within specific conditions, i.e. for some versions of a computer tutor but not others, and concludes that some student types "fit" with some computer tutor types with respect to maximizing learning.

We found no significant correlations for any user satisfaction metric over all students; this is likely because the four conditions patterned quite differently on these correlations. *Empirical* had no significant correlations. *Normal* had one trend for a negative correlation: SQ15 (R=-0.382, p=0.096), suggesting that students who perceived they had a harder time using the system actually learned more. *Random* had one trend for a positive correlation: SQ1 (R= 0.401, p = 0.089), suggesting that students who perceived they had an easier time learning from the system actually did learn more. Finally, *Basic* had one significant and one trend for a positive correlation: SQ7 (R= 0.482, p = 0.037), suggesting that students who perceived they were able to better concentrate with the system actually learned more, and SQ12 (R=0.432, p=0.065), suggesting that students who perceived that the system decreased their uncertainty actually learned more. This last result suggests that *basic* affect-adaptive ITSPOKE is "working": even if it is not the most preferred system overall, it is decreasing uncertainty while increasing learning.

4. Related Work

Other computer dialogue tutor research has evaluated and/or is developing adaptations to student affect. Much of this work bases system responses on human tutor dialogue act responses. Examples include researchers focused on feedback responses to affect. In [16] positive feedback responses were developed based on a frequency analysis of human tutor responses, including praising acknowledgments after CU answers, and were implemented in a spoken Memory Game computer tutor. Students rated the system that used the positive feedback more highly than a non-adaptive version. Similarly, in [17], a human Wizard provided positive feedback in a spoken Reading Tutor, after detecting student affect including uncertainty. The scaffolding resulted in increased student persistence as compared to a non-adaptive version. However, no other evaluation to our knowledge has shown a positive impact on student learning or learning efficiency.

Other examples include researchers focused on more contentful responses to affect. For example, [3] used a frequency analysis to extract two human tutor responses to CU and IU answers from a human tutoring corpus, then implemented and evaluated them in the SCoT-DC tutor. These adaptations increased learning when used after all correct and incorrect answers, but not when used only after uncertain answers. [2] present a machine learning analysis aimed at learning system responses from multiple human tutors' responses to student affect including uncertainty. However, their analysis suggests that because human tutors have different styles and skill levels, studying multiple human tutors does not necessarily yield consistent generalizations. [7] also study how expert human tutors vary their dialogue acts after student affect including uncertainty.

5. Conclusions and Current Directions

This study shows that adapting to student uncertainty during wizarded computer tutoring can improve learning efficiency and user satisfaction. The *basic* adaptive system showed significant improvement on learning efficiency and a positive correlation between learning and user perception of decreased uncertainty, while our *empirical* adaptive system showed a trend for improvement on user perception of tutor response quality. We hypothesize that our adaptive systems did not outperform the *random* control system for two reasons: first, the *random* system adapted to some CU answers, thus diminishing the difference with the adaptive systems; second, it adapted to CnonU answers, which may also benefit performance by increasing certainty.

Our results provide evidence that our theory-based *basic* adaptation is more effective than our human-based *empirical* adaptation. It may be that different behaviors are optimally effective in computer and human tutors. However, we do not want to conclude that empirical human tutor-based adaptations are less effective in general, for two reasons. First, our *empirical* adaptation included both feedback variation and dialogue act variation; it may be that these two components have different effectiveness. Second, our *empirical* adaptation was derived from statistical generalizations about human tutor responses to uncertainty, but not necessarily *effective* responses. In future work we will investigate other approaches that enable us to select human tutor responses that optimize learning, such as reinforcement learning or dialogue act-learning correlations.

We are currently conducting a fully automated version of this experiment, where ITSPOKE performs speech recognition, language understanding, and uncertainty detection. This experiment uses only the *basic* adaptive system (in the experimental condition) and uses the same two control conditions, except that the *random* control system adapts randomly only to CnonU answers.

Acknowledgments

NSF (#0631930) supports this work. We thank the ITSPOKE Group, Bob Hausmann, and the reviewers for helpful comments.

References

[1] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[2] Kaska Porayska-Pomsta, Manolis Mavrikis, and Helen Pain. Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18:125–173, 2008.

[3] H. Pon-Barry, K. Schultz, E. Owen Bratt, B. Clark, and S. Peters. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16:171–194, 2006.

[4] K. VanLehn, S. Siler, and C. Murray. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249, 2003.

[5] K. Forbes-Riley, M. Rotaru, and D. Litman. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*, 18(1-2):11–43, February 2008.

[6] P. Ekman and W. V. Friesen. *The facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, Palo Alto, 1978.

[7] B. Lehman, M. Matthews, S. D'Mello, and N. Person. What are you feeling? Investigating student affective states during expert human tutoring sessions. In *Intelligent Tutoring Systems Conference (ITS)*, pages 50–59, Montreal, Canada, June 2008.

[8] Sidney K. D'Mello, Scotty D. Craig, Amy Witherspoon, Bethany McDaniel, and Arthur Graesser. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18:45–80, 2008.

[9] H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of Interspeech*, pages 797–800, Pittsburgh, PA, 2006.

[10] K. Forbes-Riley and D. Litman. Designing and evaluating an uncertainty-adaptive spoken dialogue tutoring system. 2009. Under review.

[11] K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems*, 2002.

[12] K. Forbes-Riley, D. Litman, and M. Rotaru. Responding to student uncertainty during computer tutoring: A preliminary evaluation. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS)*, Montreal, Canada, June 2008.

[13] M. Ringenberg and K. VanLehn. Scaffolding problem solving with annotated worked-out examples to promote deep learning. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, pages 624–634, 2006.

[14] K. Moreno, B. Klettke, K. Nibbaragandla, and A. Graesser. Perceived characteristics and pedagogical efficacy of animated conversational agents. In *Proceedings of the Intelligent Tutoring Systems Conference (ITS)*, Biarritz, France, 2002.

[15] Mihai Rotaru. *Applications of Discourse Structure for Spoken Dialogue Systems*. PhD thesis, University of Pittsburgh, Pittsburgh, PA, 2008.

[16] W. Tsukahara and N. Ward. Responding to subtle, fleeting changes in the user's internal state. In *Proc. SIG-CHI on Human factors in computing systems*, pages 77–84, Seattle, WA, 2001. ACM.

[17] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. In *Proceedings of Intelligent Tutoring Systems Conference Workshop on Empirical Methods for Tutorial Dialogue Systems*, 2002.