Improving (Meta)cognitive Tutoring by Detecting and Responding to Uncertainty

Diane Litman and **Kate Forbes-Riley**
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

We hypothesize that enhancing computer tutors to respond to student uncertainty over and above correctness is one method for increasing both student learning and self-monitoring abilities. We explore this hypothesis using data from an experiment with a wizarded spoken tutorial dialogue system, where tutor responses to uncertain and/or incorrect student answers were manipulated. Our results suggest that monitoring and responding to student uncertainty has the potential to improve both cognitive and metacognitive student abilities.

Introduction

Speech and language researchers have shown that speaker uncertainty is associated with linguistic signals (Liscombe, Venditti, and Hirschberg 2005; Nicholas, Rotaru, and Litman 2006; Dijkstra, Krahmer, and Swerts 2006; Pon-Barry 2008), while tutoring researchers have hypothesized that tutors use such signals to detect and address student uncertainty in order to improve performance metrics including student learning, persistence, and system usability (Tsukahara and Ward 2001; Aist et al. 2002; Litman et al. 2009). For example, VanLehn et al. (2003) propose that both student uncertainty and incorrectness signal "learning impasses", i.e. student learning opportunities. While correlational studies have shown a link between student uncertainty and learning in tutorial dialogue (Craig et al. 2004; Forbes-Riley, Rotaru, and Litman 2008), few controlled experiments have investigated whether responding to student impasses involving uncertainty improves learning (e.g. (Pon-Barry et al. 2006)); most computer dialogue tutors respond based only on student correctness.

In prior work, we experimentally compared learning gains and efficiency across four versions of a wizarded spoken tutorial dialogue system that differed in whether and how it adapted to student uncertainty (Forbes-Riley and Litman 2009a). Uncertainty and correctness were manually annotated in real-time by a human "Wizard". In two experimental conditions, the system provided additional knowledge at places of uncertainty. In two control conditions, the system either did not provide this knowledge, or provided such knowledge randomly. Our results suggested that responding to student uncertainty, over and above correctness, did indeed lead to performance improvements along cognitive dimensions. Here we turn our attention to student metacognition. In particular, we present empirical evidence suggesting that by responding to student uncertainty in new ways, our computer tutor has the potential to further improve both cognitive and metacognitive performance.

System and Data

This research uses a corpus of 405 dialogues between 81 students and wizarded versions of ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system), a speech-enabled version of the Why2-Atlas qualitative physics tutor (VanLehn et al. 2002). The corpus was collected in our prior experiment evaluating the utility of enhancing ITSPOKE to respond to impasses involving student uncertainty, over and above correctness (Forbes-Riley and Litman 2009a).

Motivated by research that views uncertainty as well as incorrectness as signals of "learning impasses" (VanLehn, Siler, and Murray 2003), ITSPOKE was modified to associate one of four impasse states with every student answer. The four impasse states correspond to all possible combinations of (binary) uncertainty (uncertain (**U**), nonuncertain (**nonU**)[1]) and correctness (incorrect (**I**), correct (**C**)), as shown in Figure 1.

Nominal State:	InonU	IU	CU	CnonU
Scalar State:	3	2	1	0
Severity Rank:	most	less	least	none

Figure 1: Different Impasse State Severities

The incorrectness component of each state reflects the actual accuracy of the student's answer, while the uncertainty component reflects the tutor's perception of the student's awareness of this accuracy. The scalar ranking of impasse states in terms of severity combines these two components and will be discussed below.

The experimental procedure for collecting the corpus used a semi-automatic version of ITSPOKE, where a human

[1] A 'nonuncertain' answer may be certain or neutral.

ITSPOKE1: If the force of the man weren't greater than the force of gravity, the pumpkin wouldn't be tossed up into the air. But we know that it is. So the direction of the net force must be upwards. Does the pumpkin accelerate while the man is tossing it? If yes, please specify the direction of the acceleration.

STUDENT1: um... up. *[CU]*

ITSPOKE2: Fine. *[FEEDBACK]*

Figure 2: Example of *Normal* ITSPOKE Response to **CU** Answers

ITSPOKE1: If the force of the man weren't greater than the force of gravity, the pumpkin wouldn't be tossed up into the air. But we know that it is. So the direction of the net force must be upwards. Does the pumpkin accelerate while the man is tossing it? If yes, please specify the direction of the acceleration.

STUDENT1: Yes, uh vertically ... up, well, yeah, vertically up. *[CU]*

ITSPOKE2: Right. *[FEEDBACK]* To analyze the pumpkin's acceleration we will use Newton's Second law. What's the definition of this law? *[REMEDIATION SUBDIALOGUE]*

Figure 3: Example of *Basic* Uncertainty Adaptation (for **CU** Answers only)

"wizard" performed speech recognition, as well as correctness and uncertainty annotation. That is, each student turn was annotated in real-time by the wizard during the experiment, using binary student *Uncertainty* (**U**, **nonU**) and *Correctness* (**C**, **I**) tags[2]. Subjects who had never taken college physics: read a short physics text; took a multiple-choice pretest; worked 5 problems (i.e. engaged in 5 dialogues) with ITSPOKE; took a survey; took an isomorphic posttest.

The experiment had two control and two experimental conditions. The *normal* control condition remediated only incorrectness impasses (**InonU, IU**), as in the original ITSPOKE. An example dialogue excerpt from this condition is shown in Figure 2. As shown, **ITSPOKE2** provides correctness feedback for the **CU** answer, and ignores the uncertainty.

In contrast, the two experimental conditions remediated both uncertainty and incorrectness impasses (**InonU, IU, CU**), but each used a different approach. The *basic* experimental condition used the same remediation for all impasse types, with only feedback phrases varying based on answer correctness (e.g. "That's right" versus "That's wrong"). An example dialogue excerpt is shown in Figure 3. As shown, **ITSPOKE2** provides correctness feedback for the **CU** answer, then responds to the uncertainty by providing the same remediation subdialogue (i.e., a series of additional questions) that would have been provided if the student answer were incorrect. Only the first question in this remediation subdialogue is shown. Note that **IU and InonU** answers already receive this remediation subdialogue (because they are incorrect), therefore the *basic* uncertainty adaptation impacts only **CU** answers.

In contrast to the *basic* experimental condition, the *empirical* experimental condition used different dialogue act presentations of the incorrect answer content (e.g. remediation subdialogue questions versus "bottom out" statements) *and* different feedback phrases (e.g. "That's exactly right, but you seem unsure" for **CU** versus "Great try, but I think you know there is a mistake in your answer" for **IU**) to respond to each impasse type (**CU, IU, InonU**). Example dialogue excerpts for each impasse type are shown in Figure 4. As shown, in each case **ITSPOKE2** provides feedback responding to the specific impasse type of the answer, then provides the same additional content responding to the uncertainty of the answer, but the dialogue act used to present this additional content depends on the specific impasse type of the answer.[3]

Finally, the *random* control condition treated a percentage of random correct answers as incorrect, to control for the additional content in the experimental conditions. The motivation for and further details of each experimental condition are discussed in detail elsewhere (Forbes-Riley and Litman 2009a).

Metacognitive Measures

In this section we introduce several ways of combining the corpus uncertainty and correctness annotations into single quantitative performance measures. Note that all measures are computed on a per student basis (over all five dialogues).

Our first measure is based on a ranking of impasses by severity. In particular, we first associate a scalar **impasse severity** value with each student answer in our corpus, based on our wizard's correctness and uncertainty annotations. We then compute an average impasse severity per student, according to whether the impasses were due to uncertainty, incorrectness, or both. Our severity values were proposed in our earlier work (Forbes-Riley, Litman, and Rotaru 2008) and are shown in Figure 1. According to our ranking, the most severe type of impasse (severity 3) occurs when a student is incorrect but not aware of it. States of severity 2 and 1 are of increasingly lesser severity: the student is incorrect but aware that s/he might be, and the student is correct but uncertain about it, respectively. Finally, no impasse exists when a student is correct and not uncertain about it (severity 0). These severity rankings reflect our belief that to resolve an impasse, a student must first perceive that it exists. Incorrectness simply indicates that the student has reached an impasse, while uncertainty - in a correct or incorrect answer

[2] This wizard displayed interannotator agreement of 0.85 and 0.62 Kappa on labeling correctness and uncertainty, respectively, in prior studies (Forbes-Riley and Litman 2008).

[3] The dialogue act variations were developed based on analysis of human tutor responses to uncertainty in a human tutoring corpus (see (Forbes-Riley and Litman 2009a) for further details)

ITSPOKE1: If the force of the man weren't greater than the force of gravity, the pumpkin wouldn't be tossed up into the air. But we know that it is. So the direction of the net force must be upwards. Does the pumpkin accelerate while the man is tossing it? If yes, please specify the direction of the acceleration.

STUDENT1: It accelerates upwards?? *[CU]*

ITSPOKE2: Great answer, but you don't sound very certain so let's go over it. *[FEEDBACK]* The pumpkin does have an acceleration and it points vertically up. We would use Newton's Second law to prove this (Net Force = Mass * Acceleration). Since the net force on the pumpkin is non-zero and points up, this law tells us its acceleration must also be non-zero and point up. *[NEW BOTTOM OUT]*

STUDENT1: Vertically?? *[IU]*

ITSPOKE2: I'm sorry. It sounds like you already know there's an error. Let's work it out. *[FEEDBACK]* The pumpkin does have an acceleration and it points vertically up. Let's work through the proof of this step by step.*[NEW BOTTOM OUT]* To analyze the pumpkin's acceleration we will use Newton's Second law. What's the definition of this law? *[REMEDIATION SUBDIALOGUE]*

STUDENT1: Yes, negative 9.8 meters per second squared. *[InonU]*

ITSPOKE2: Good try, but there's a mistake in your answer that we need to work through. *[FEEDBACK]* To analyze the pumpkin's acceleration we will use Newton's Second law. What's the definition of this law? *[REMEDIATION SUBDIALOGUE]*

Figure 4: Example of *Empirical* Uncertainty Adaptation for **CU, IU, and InonU** Answers

- indicates that the student perceives s/he has reached an impasse.

The rest of our measures are taken from the metacognitive performance literature. The knowledge monitoring accuracy measure that we use is the Harmann coefficient **(HC)** (Nietfeld, Enders, and Schraw 2006).[4] This measure has previously been used to measure the monitoring accuracy of one's own knowledge ("Feeling of Knowing" (FOK)), which is closely related to uncertainty. Psycholinguistics research has shown that speakers display FOK in conversation using linguistic cues (Smith and Clark 1993), and that listeners can use the same cues to monitor the FOK of someone else ("Feeling of Another's Knowing" (FOAK)) (Brennan and Williams 1995). High and low FOK/FOAK judgments have also been associated with speaker certainty and uncertainty, respectively (Dijkstra, Krahmer, and Swerts 2006).

	Correct	Incorrect
Nonuncertain	CnonU	InonU
Uncertain	CU	IU

Figure 5: Measuring Student Metacognitive Performance

HC is computed from our wizard's correctness and uncertainty annotations as follows:

$$HC = \frac{(CnonU+IU)-(InonU+CU)}{(CnonU+IU)+(InonU+CU)}$$

HC ranges in value from -1 (no monitoring accuracy) to 1 (perfect monitoring accuracy).

To illustrate the reasoning behind HC and the other metacognitive performance measures used in this paper, consider an FOK-type experimental paradigm, where subjects 1) respond to a set of general knowledge questions, 2) take a survey, judging whether or not[5] they think they would be uncertain about the answer to each question in a multiple choice test, and 3) take such a multiple choice test. In FOAK-type paradigms such as ours, the *wizard* annotates the correctness and uncertainty for each student answer. As shown in Figure 5, such FOK or FOAK data can be summarized in an array where each cell represents a mutually exclusive option: the row labels represent the possible uncertainty judgments (Nonuncertain or Uncertain), while the columns represent the possible correctness results of the multiple choice test (Correct or Incorrect). Given such an array, various relationships between the correctness of answers, and the judged uncertainty of the answers, can then be computed.

Following (Saadawi et al. 2009), who investigate the role of immediate feedback and other metacognitive scaffolds in a medical tutoring system, we additionally measure metacognitive performance in terms of **bias** and **discrimination** (Kelemen, Frost, and Weaver 2000). As with HC, we compute these measures using our wizard's correctness and uncertainty annotations. Bias scores greater than and less than zero indicate overconfidence and underconfidence, respectively, with zero indicating best metacognitive performance:

$$bias = \frac{CnonU+InonU}{CnonU+InonU+CU+IU} - \frac{CnonU+CU}{CnonU+InonU+CU+IU}$$

In contrast, discrimination scores greater than zero indicate higher metacognitive performance, in terms of certainty for correct responses and uncertainty for incorrect responses:

$$discrimination = \frac{CnonU}{CnonU+CU} - \frac{InonU}{InonU+IU}$$

To illustrate the computation of our metacognitive performance metrics, suppose the annotated dialogue excerpt in Figure 4 represented our entire dataset (from a single student). Then we would have the following values for our metrics for that student:

[4] While the Gamma measure is often also used, there is a lack of consensus regarding the relative benefits of Gamma versus HC (Nietfeld, Enders, and Schraw 2006), and we have found HC to be more predictive for our corpus.

[5] Likert scale rating schemes are also possible.

$$HC = \frac{(0+1)-(1+1)}{(0+1)+(1+1)} = -\frac{1}{3}$$

$$bias = \frac{0+1}{0+1+1+1} - \frac{0+1}{0+1+1+1} = \frac{1}{3} - \frac{1}{3} = 0$$

$$discrimination = \frac{0}{0+1} - \frac{1}{1+1} = \frac{0}{1} - \frac{1}{2} = -\frac{1}{2}$$

Results

In this section we investigate whether the measures introduced in the previous section differ across our experimental conditions, and/or predict student learning gains. We first 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*. Even though our experiment was designed to only impact learning gain, we hypothesized that the experimental conditions might still reduce impasse severity: by responding contingently to uncertainty the tutor resolved more impasse types. For similar reasons, we hypothesized that the experimental conditions might also improve student accuracy in monitoring their own uncertainty (i.e., FOK), particularly in the empirical condition where the tutor's feeling of the student's uncertainty (i.e., FOAK) was explicitly stated. Our HC metric measures inferred (rather than actual) student self-monitoring accuracy (because it was derived from our wizard's uncertainty labels, rather than student judgments of their own uncertainty). We had similar hypotheses for bias and discrimination.

The "Means" columns in Table 1 show the means per condition. As predicted, both experimental conditions had lower average impasse severity than *random*, and *random* was lower than *normal*. While a one-way ANOVA with post-hoc Tukey showed no statistically significant differences or trends among these means (p = .19), paired contrasts showed trends for individual differences between *random* and *normal* (p = .10), *basic* and *normal* (p = .06), and between *empirical* and *normal* (p = .08). With respect to both inferred self-monitoring accuracy (HC) and bias, the ANOVAS showed no statistically significant differences or trends across conditions. However, for HC the paired contrasts showed a trend for differences between *basic* and *normal* (p = .06), and *random* and *normal* (p = .06), in the predicted directions. With respect to discrimination, the ANOVA indicated a trend for a difference among the means (p = .09), with paired contrasts showing significant differences between *basic* and *empirical* (p = .04), and between *random* and *empirical* (p = .02); note, however, that contrary to our predictions, discrimination was lowest in the empirical condition.

Although we only find weak support for differences in metacognitive performance between conditions, we still hypothesize that lower impasse severities, higher self-monitoring accuracies, less bias, and better discrimination are better for students from a learning perspective. To support this view, we computed a partial Pearson's correlation over all 81 students between each measure and posttest score, controlled for pretest score to measure learning gain. The last two columns in Table 1 show the Pearson's Correlation Coefficient (R) of the partial correlation, and the significance of the correlation (p). As predicted, average impasse severity is significantly negatively correlated with learning, while inferred self-monitoring accuracy (HC) and discrimination are significantly positively correlated with learning. There is also a trend for bias to be negatively correlated with learning, suggesting that underconfidence is better than overconfidence.

Discussion

We presented an analysis of student metacognitive performance using data from a wizarded dialogue tutor that adapts to student uncertainty. The performance measures examined include several traditional measures of metacognitive performance, as well as a new learning impasse severity measure we derived from a theory of uncertainty and incorrectness as learning impasses. While our prior work demonstrated that remediating after uncertainty impasses improves learning (Forbes-Riley and Litman 2009a), the results in Table 1 suggest that further investigation into better ways of remediating student uncertainty holds promise for further improving student cognitive as well as metacognitive abilities.

Our correlations show that both **average impasse severity** and (tutor perception of) **self-monitoring accuracy** and **discrimination** significantly predict student learning (negatively, positively, and positively respectively). Although correlation does not imply causality, our findings motivate future modifications of our system to increase student learning. For example, we plan to develop remediations that are better optimized for each impasse type, particularly for impasses with the highest severity. We also plan to enhance our tutor to not only remediate domain content after impasses (as in the current experiment), but to also remediate inferred student knowledge monitoring abilities.

While our ANOVAS show that **impasse severity** doesn't differ significantly across conditions, the means are consistent with our predictions, and there are statistical pairwise trends suggesting improvement between all conditions and *normal* (the original system). We also see similar results for *basic* and *random* compared to *normal* with respect to inferred self-monitoring accuracy (**HC**). These are promising findings, as our current interventions were designed to improve only student correctness on the posttest, not to reduce impasse severity or increase monitoring accuracy. In the future we would like to enhance our interventions to directly target student knowledge monitoring, and to better measure such improvements by incorporating FOK ratings into our testing. There is increasing interest in using intelligent tutoring systems to teach metacognition, and we plan to build on this literature (e.g. (Aleven and Roll 2007; Roll and Aleven 2008; Saadawi et al. 2009)).

We found it surprising that neither experimental condition outperformed *random*. We hypothesize that this is because CU impasses are sometimes adapted to in *random*; in future versions of our system, we will only randomly remediate after only CnonU answers (non-impasse states). We also plan to revisit how we designed our *empirical* experimen-

Measure	Means				Correlation (81)	
	normal (21)	random (20)	basic (20)	empirical (20)	R	p
Average Impasse Severity	.73	.60	.59	.59	-.56	.00
Monitoring Accuracy	.52	.62	.62	.58	.42	.00
Bias	-.02	-.01	-.03	-.01	-.21	.06
Discrimination	.41	.48	.46	.34	.32	.00

Table 1: Means Across Experimental Conditions, and Partial Correlations with Posttest, for Impasse Severity, Monitoring Accuracy, Bias, and Discrimination

tal condition, as it did not yield the expected performance improvements.

Finally, we plan to replicate the analyses in this study, using a dialogue corpus that was recently collected using a fully automated version of ITSPOKE that detects and adapts to student uncertainty. We also recently found interactions between learning and user classes based on user domain expertise and gender in the wizarded corpus (Forbes-Riley and Litman 2009b); we will investigate whether the interactions with these classes extend to the student metacognitive metrics discussed in this paper.

Acknowledgments

This work is supported by NSF Award 0631930.

References

Aist, G.; Kort, B.; Reilly, R.; Mostow, J.; and Picard, R. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proc. Intelligent Tutoring Systems Workshop on Empirical Methods for Tutorial Dialogue Systems*.

Aleven, V., and Roll, I., eds. 2007. *AIED Workshop on Metacognition and Self-Regulated Learning in Intelligent Tutoring Systems*.

Brennan, S. E., and Williams, M. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*.

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

Dijkstra, C.; Krahmer, E.; and Swerts, M. 2006. Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In *Proc. Speech Prosody*.

Forbes-Riley, K., and Litman, D. J. 2008. Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In Dybkjaer, L., and Minker, W., eds., *Recent Trends in Discourse and Dialogue*. Springer.

Forbes-Riley, K., and Litman, D. 2009a. Adapting to student uncertainty improves tutoring dialogues. In *Proc. Intl. Conf. on Artificial Intelligence in Education*.

Forbes-Riley, K., and Litman, D. 2009b. A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In *Proceedings of Interspeech*.

Forbes-Riley, K.; Litman, D.; and Rotaru, M. 2008. Responding to student uncertainty during computer tutoring: A preliminary evaluation. In *Proc. Intl. Conf. on Intelligent Tutoring Systems*.

Forbes-Riley, K.; Rotaru, M.; and Litman, D. 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*.

Kelemen, W. L.; Frost, P. J.; and Weaver, C. A. 2000. Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory and Cognition* 28:92–107.

Liscombe, J.; Venditti, J.; and Hirschberg, J. 2005. Detecting certainness in spoken tutorial dialogues. In *Proc. Interspeech*.

Litman, D.; Moore, J.; Dzikovska, M.; and Farrow, E. 2009. Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In *Proc. Intl. Conf. on Artificial Intelligence in Education*.

Nicholas, G.; Rotaru, M.; and Litman, D. J. 2006. Exploiting word-level features for emotion prediction. In *Proc. IEEE/ACL Workshop on Spoken Language Technology*.

Nietfeld, J. L.; Enders, C. K.; and Schraw, G. 2006. A monte carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement*.

Pon-Barry, H.; Schultz, K.; Bratt, E. O.; Clark, B.; and Peters, S. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education* 16.

Pon-Barry, H. 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In *Proc. Interspeech*.

Roll, I., and Aleven, V., eds. 2008. *ITS Workshop on Meta-Cognition and Self-Regulated Learning in Educational Technologies*.

Saadawi, G. M. E.; Azevedo, R.; Castine, M.; Payne, V.; Medvedeva, O.; Tseytlin, E.; Legowski, E.; azen Jukic, D.; and Crowley, R. S. 2009. Factors affecting feeling-of-knowing in a medical intelligent tutoring system: the role of immediate feedback as a metacognitive scaffold. *Adv in Helth Sci Educ*.

Smith, V. L., and Clark, H. H. 1993. On the course of answering questions. *Journal of Memory and Language*.

Tsukahara, W., and Ward, N. 2001. Responding to subtle, fleeting changes in the user's internal state. In *Proc. SIGCHI on Human factors in computing systems*.

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

VanLehn, K.; Siler, S.; and Murray, C. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction* 21(3).