25 Towards Improving (Meta)cognition by Adapting to Student Uncertainty in Tutorial Dialogue

Diane Litman and Kate Forbes-Riley

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 test this hypothesis using spoken data from both wizarded and fully-automated versions of a spoken tutorial dialogue system, where tutor responses to uncertain and/or incorrect student answers were manipulated. Although we find no significant improvement in metacognitive metrics (computed using speech and language information) when responding to uncertainty and incorrectness as compared to when responding only to incorrectness, we find that some metacognitive metrics significantly correlate with student learning. 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 (Dijkstra, Krahmer, & Swerts, 2006, Liscombe, Venditti, & Hirschberg, 2005, Nicholas, Rotaru, & Litman, 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 (Aist, Kort, Reilly, Mostow, & Picard, 2002, Litman, Moore, Dzikovska, & Farrow, 2009, Tsukahara & Ward, 2001). 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 learning and student uncertainty as well as the related notion of confusion in tutorial dialogue (Craig, Graesser, Sullins, & Gholson, 2004, Forbes-Riley, Rotaru, & Litman, 2008b), few controlled experiments have investigated whether responding to student impasses involving uncertainty improves learning, and those that did yielded overall null results (e.g., (Pon-Barry et al., 2006)). To date, most computer dialogue tutors respond based only on student correctness.

D. Litman (✉) • K. Forbes-Riley
Learning Research and Development Center,
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: litman@cs.pitt.edu

R. Azevedo and V. Aleven (eds.), *International Handbook of Metacognition and Learning Technologies*,
Springer International Handbooks of Education 26, DOI 10.1007/978-1-4419-5546-3_25,

In prior work, we experimentally compared learning gains across versions of a spoken tutorial dialogue system that differed in whether and how they adapted to student uncertainty. In our experimental conditions, the system provided additional knowledge at places of uncertainty; in the control conditions, the system either did not provide this knowledge, or provided such knowledge randomly. In a first experiment we used a wizarded form of our system, where uncertainty and correctness were manually annotated in real time by a human "wizard" (Forbes-Riley & Litman, 2009a, 2011b). Our results demonstrated that responding to student uncertainty, over and above correctness, did indeed lead to performance improvements along cognitive dimensions. In a subsequent experiment we used a fully automated version of our system, where uncertainty in each student turn was (noisily) detected using acoustic-prosodic and lexical features extracted from the speech signal, as well as dialogue features. Our results were again that enhancing our system to respond to uncertainty yielded higher student learning gains than non-adaptive control systems, but here the difference was only significant for a subset of students after we controlled for the proportion of additional tutoring content received during the tutoring interaction. In particular, students who received the adaptation learned significantly more than students in a control condition who randomly received an equal proportion of additional tutoring content. Based on system error analyses we concluded that the uncertainty adaptation had only a small effect on learning in the fully automated system because the system did not automatically recognize student uncertainty often enough and thus did not give the adaptation often enough (see Forbes-Riley & Litman (2011a) for further details).

In this chapter we turn our attention to student metacognition. First, we show how to construct measures of student metacognitive performance (e.g., monitoring accuracy, bias, discrimination) using the manually and automatically created tutor annotations of student uncertainty and correctness available from our prior wizarded and fully automated experiments, respectively. Next, we examine whether our prior tutor adaptations to student uncertainty—which have already been shown to improve cognition—can also improve metacognition. Finally, we examine whether our measures of metacognitive performance are correlated with our measures of cognitive performance (i.e., learning gain), and whether such correlations are robust to the noise introduced by speech and language processing techniques. Analyses of the data from both our wizarded and fully automated experiments demonstrate that by responding to student uncertainty in new ways, tutorial dialogue systems have the potential to further improve both cognitive and metacognitive performance.

Systems and Data

This research uses corpora of dialogues (see Figs. 25.2–25.3 for examples) between students and both ITSPOKE-WOZ and ITSPOKE-AUTO, wizarded and fully automated versions of ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system), respectively. ITSPOKE in turn is a speech-enabled version of the Why2-Atlas qualitative physics tutor (VanLehn et al., 2002), which asks "why-type" questions relating to Newtonian physics.[1] The corpora were collected in our prior experiments evaluating the utility of enhancing ITSPOKE to respond to impasses involving student uncertainty over and above correctness, in wizarded (Forbes-Riley & Litman, 2009a, 2011b) and fully automated (Forbes-Riley & Litman, 2011a) conditions. The target audience for ITSPOKE are novices, i.e., college students who have never taken college-level physics.

The conceptual framework of our work is based on the theory of learning impasses. Motivated by research that views uncertainty as well as

[1] The version of ITSPOKE used here differs from the original ITSPOKE and Why2-Atlas in that the system has been reimplemented using the TuTalk tools for authoring tutorial dialogue systems (Jordan, Hall, Ringenberg, Cui, & Rosé, 2007) and does not include the essay writing component of Why2-Atlas. As will be discussed, several versions of ITSPOKE used in the experiments reported here (in particular, in the experimental but not in the control conditions) have in addition been enhanced to detect and adapt to student uncertainty.

Fig. 25.1 Different Impasse State Severities

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

incorrectness as signals of "learning impasses" (VanLehn et al., 2003), i.e., opportunities for the student to learn the material that he/she is incorrect or uncertain about, the original version of 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 student *uncertainty* (uncertain (**U**), nonuncertain (**nonU**)[2]) and *correctness* (incorrect (**I**), correct (**C**)), as shown in Fig. 25.1.[3]

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. While the original ITSPOKE only remediated incorrectness impasses (InonU and IU states), our uncertainty-adaptive ITSPOKE also remediates all uncertainty impasses (CU states – note that IU impasses were already remediated in the original non-adaptive system). Impasse theory is similar to cognitive disequilibrium theory (Craig, Graesser, Sullins, & Gholson, 2004), which predicts that confusion is likely to occur during cognitive disequilibrium, and that trying to restore equilibrium will lead to learning gains.

ITSPOKE-WOZ

The ITSPOKE-WOZ corpus consists of 405 dialogues between 81 students and ITSPOKE-WOZ, a semi-automatic version of ITSPOKE where a human "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, producing the binary student uncertainty and correctness tags.[4] Using a wizard allowed us to examine the impact of adapting to uncertainty impasses in upper-bound tutorial dialogue conditions, that is, without the errors introduced by using automated speech and language processing components. In both human and automatic detection of uncertainty, common indicators of student uncertainty include lexical hedges (e.g.,"I think"), pitch features (e.g., rising intonation), temporal features (e.g., pausing), and energy features (e.g,. soft-spokeness). Such features have been validated both through interannotator agreement studies and automatic detection studies (c.f., Forbes-Riley & Litman 2011a).

The experimental procedure for collecting the corpus was as follows: subjects who had never taken college physics read a short physics text, took a multiple-choice pretest, worked five problems[5] (i.e., engaged in five dialogues) with ITSPOKE, took a survey[6], and took an isomorphic posttest.

The experiment had two control and two experimental conditions. Complete details about these conditions are provided elsewhere (Forbes-Riley & Litman, 2011b); here we briefly outline the differences between them.

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

[3] For example, the 6,561 student turns in the ITSPOKE-WOZ corpus are distributed among these states as follows: 650 InonU (10%), 764 IU (12%), 727 CU (11%), and 4,420 CnonU (67%).

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

[5] For example, the problem referred to in the dialogue excerpts in Figs. 25.2–25.3 is: "Suppose a man is running in a horizontal line at a constant velocity. He tosses a pumpkin vertically up while he is running. Where will the pumpkin land relative to the man? Assume that air resistance is negligible."

[6] The survey is not used in this paper; see (Forbes-Riley & Litman, 2009a) for the survey usage.

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]*

Fig. 25.2 Example of *nonAdapt* 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]*

Fig. 25.3 Example of *Simple* Uncertainty Adaptation (for **CU** Answers only)

The *nonAdapt* control condition remediated only incorrectness impasses (**InonU, IU**), as in the original ITSPOKE. An example dialogue excerpt from this condition is shown in Fig. 25.2 . As shown, **ITSPOKE2** provides feedback indicating the correctness of 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 *Simple* 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 Fig. 25.3 . As shown, **ITSPOKE2** provides feedback indicating the correctness of the **CU** answer and 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 *Simple* uncertainty adaptation impacts only **CU** answers.

In contrast to the *Simple* experimental condition, the *Complex* 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 Fig. 25.4 . As shown, in each case, **ITSPOKE2** provides feedback responding to the specific impasse type of the answer and then provides the same additional content responding to the impasse, but the dialogue act used to present

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]*

Fig. 25.4 Example of *Complex* Uncertainty Adaptation for **CU, IU, and InonU** Answers

this additional content depends on the specific impasse type of the answer.[7]

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 & Litman, 2009a, 2011b).

ITSPOKE-AUTO

The ITSPOKE-AUTO corpus consists of 360 dialogues between 72 students and ITSPOKE-AUTO, a fully automated version of ITSPOKE in which speech recognition as well as correctness and uncertainty annotation were automatically performed by speech and language processing components. Student speech was digitized from microphone input and sent to the Sphinx2 speech recognizer (Huang et al., 1993), whose stochastic language models were trained on the ITSPOKE-WOZ corpus and prior ITSPOKE corpora. Correctness was automatically labeled on the

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

speech recognition output using the TuTalk semantic analyzer (Jordan et al., 2007), which was trained on the ITSPOKE-WOZ corpus. Uncertainty was automatically labeled on the speech recognition output using an uncertainty model built with WEKA software (Witten & Frank, 1999) from features of the student speech and dialogue context, including lexical, pitch, temporal, and energy features as well as tutor question and gender. The uncertainty model is a logistic regression equation that was trained on the ITSPOKE-WOZ corpus, where the wizard's labels were the ground truth labels. The most important predictors of student uncertainty in the model were pitch and lexical features of the student's current turn, as well as the type of tutor question in the preceding turn.

The ITSPOKE-AUTO corpus was collected using the procedure from the ITSPOKE-WOZ experiment, although the experimental conditions were changed in two ways. First, the *Complex* experimental condition was removed. We removed this condition as only *Simple* yielded learning improvements for ITSPOKE-WOZ (Forbes-Riley & Litman, 2009a, 2011b). Second, *Random* was changed so that ITSPOKE-AUTO randomly remediated after only CnonU answers (non-impasse states). We changed this condition because in ITSPOKE-WOZ neither wizarded experimental condition outperformed *Random* (Forbes-Riley & Litman, 2009a, 2011b); we hypothesized this was because CU impasses were sometimes adapted to in *Random*. Full details of the ITSPOKE-AUTO system, including a performance analysis of the speech and language processing components and their impact on the learning results, are presented elsewhere (Forbes-Riley & Litman, 2010, 2011a).

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 either our wizard's or automatically computed 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, & Rotaru, 2008a) and are shown in Fig. 25.1. 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 he/she 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—indicates that the student perceives he/she has reached an impasse.

From the standpoint of measuring metacognition, average impasse severity represents the simplest of our measures. Each impasse state reflects a current state of "self-monitoring": states 1 and 3 are currently inaccurate self-monitoring, while states 0/2 are currently perfect self-monitoring. However, the ranking of states adds a further cognitive component to the metric, by indicating how far the current self-monitoring state is from objective correctness.

The rest of our measures are taken from the metacognitive performance literature. The knowledge monitoring accuracy measure that we use is the Hamann coefficient **(HC)** (Nietfeld, Enders, & Schraw, 2006).[8] 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

[8] While the Gamma measure is often also used, there is a lack of consensus regarding the relative benefits of Gamma versus HC (Nietfeld et al., 2006), and we have found HC to be more predictive for our corpus (Litman & Forbes-Riley, 2009b).

	Correct	Incorrect
Nonuncertain	CnonU	InonU
Uncertain	CU	IU

Fig. 25.5 Measuring Student Metacognitive Performance

display FOK in conversation using linguistic cues (Smith & Clark, 1993) and that listeners can use the same cues to monitor the FOK of someone else ("feeling of another's knowing" (FOAK)) (Brennan & Williams, 1995). High and low FOK/FOAK judgments have also been associated with speaker certainty and uncertainty, respectively (Dijkstra et al., 2006).

HC measures absolute knowledge monitoring accuracy, or the accuracy with which certainty reflects correctness. HC ranges in value from -1 (no knowledge monitoring accuracy) to 1 (perfect accuracy). We compute HC from our correctness and uncertainty annotations as shown below; the numerator subtracts cases where (un)certainty is at odds with (in)correctness from cases where they correspond, while the denominator sums over all cases.

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

To illustrate the reasoning behind HC and the other metacognitive performance measures used in this paper, consider an FOK-type experimental paradigm (Smith & Clark, 1993), where subjects (1) respond to a set of general knowledge questions, (2) take a survey, judging whether or not[9] 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 *tutor* annotates the correctness and uncertainty for each student answer. As shown in Fig. 25.5 , 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, & Weaver, 2000). As with HC, we compute these measures using our tutor's correctness and uncertainty annotations.

Bias measures the overall degree to which confidence matches correctness. Bias scores greater than and less than zero indicate overconfidence and underconfidence, respectively, with zero indicating best metacognitive performance. We compute bias as shown below. The first term represents the relative proportion of confident answers (certain cases/all cases); the second represents the relative proportion of correct answers.

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

Discrimination measures the ability to discriminate performance in terms of (in)correctness. Discrimination scores greater than zero indicate higher metacognitive performance. As shown below, the first term represents the proportion of correct answers judged as certain, and the second term represents the proportion of incorrect answers judged as certain.

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

To illustrate the computation of our metacognitive performance metrics, suppose the annotated dialogue excerpt in Fig. 25.4 represented our entire dataset (from a single student). Then we would have the following values for our metrics for that student:

[9] Likert scale rating schemes are also possible.

Table 25.1 Means across ITSPOKE-WOZ experimental conditions and partial Correlations with posttest, for impasse severity, monitoring accuracy, bias, and discrimination

	Means				Correlation	
Measure	nonAdapt	Random	Simple	Complex	*R*	*p*
	(*n*=21)	(*n*=20)	(*n*=20)	(*n*=20)	(*n*=81)	
Impasse severity	0.73	0.60	0.59	0.59	−0.56	0.00
Monitoring accuracy	0.52	0.62	0.62	0.58	0.42	0.00
Bias	−0.02	−0.01	−0.03	−0.01	−0.21	0.06
Discrimination	0.41	0.48	0.46	0.34	0.32	0.00

$$\text{impasse severity} = \frac{(1+2+3)}{3} = 2$$

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

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

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

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, using the corpora from both the ITSPOKE-WOZ and ITSPOKE-AUTO experiments. We first run 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: *Complex* > *Simple* > *Random* > *nonAdapt*. 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 responded to, and thus perhaps 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 *Complex* 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 tutor's uncertainty labels, rather than student judgments of their own uncertainty). We had similar hypotheses for bias and discrimination.

Second, we compute a partial Pearson's correlation over all students between each metacognitive measure and posttest score, controlled for pretest score to measure learning gain. We hypothesized that even if we did not find any metacognitive differences between conditions, lower impasse severities, higher self-monitoring accuracies, less bias, and better discrimination would still be better for students overall, from a cognitive perspective. Our rational for this hypothesis was, simply put, that students who are more accurate in their self-monitoring know when their answers are incorrect, and thus know when to take steps to correct their errors after the system provides the correct answer and the reasoning behind it.

ITSPOKE-WOZ

The "Means" columns in Table 25.1 show the means per condition in the ITSPOKE-WOZ experiment, where each metacognitive measure was computed using the wizard's uncertainty and correctness annotations. As predicted, both experimental conditions had lower average impasse severity than *Random*, and *Random* was lower than *nonAdapt*. While a one-way ANOVA with post hoc Tukey showed no statistically significant differences or trends among these means (p = 0.19), paired contrasts showed trends for individual differences between *Random* and *nonAdapt* (p = 0.10), *Simple* and *nonAdapt* (p = 0.06), and between *Complex* and *nonAdapt* (p = 0.08). With respect to both inferred self-monitoring accuracy (HC) and bias, the ANOVAs showed no

Table 25.2 Means across ITSPOKE-AUTO experimental conditions, and partial correlations with posttest, for impasse severity, monitoring accuracy, bias, and discrimination

	Means			Correlation	
Measure	nonAdapt	Random	Simple	*R*	*p*
	(*n*=25)	(*n*=23)	(*n*=24)	(*n*=72)	
Impasse severity	0.94	0.98	0.98	−0.40	0.001
Monitoring accuracy	0.44	0.41	0.42	0.35	0.003
Bias	0.21	0.20	0.22	−0.36	0.002
Discrimination	0.19	0.20	0.19	−0.04	0.768

statistically significant differences or trends across conditions. However, for HC, the paired contrasts showed a trend for differences between *Simple* and *nonAdapt* ($p = 0.06$), and *Random* and *nonAdapt* ($p = 0.06$) in the predicted directions. With respect to discrimination, the ANOVA indicated a trend for a difference among the means ($p = 0.09$), with paired contrasts showing significant differences between *Simple* and *Complex* ($p = 0.04$), and between *Random* and *Complex* ($p = 0.02$); note, however, that contrary to our predictions, discrimination was lowest in *Complex*.

Although we only find weak support for differences in metacognitive performance between conditions, we still hypothesize that better metacognitive performance is better for students from a learning perspective. The last two columns in Table 25.1 show the Pearson's Correlation Coefficient (R) between each metacognitive measure and posttest after controlling for pretest, and the significance of the correlation (p), over all 81 students. As predicted, average impasse severity is significantly negatively correlated with learning,[10] 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.

ITSPOKE-AUTO

The "Means" columns in Table 25.2 show the means per condition in the ITSPOKE-AUTO experiment, where each metacognitive measure was computed using the automatic uncertainty and correctness annotations. The table shows that the differences were typically not in the predicted directions, although nothing was statistically significant.[11] These results thus suggest that once noise is introduced after automating speech and language processing, we no longer see even weak support for improvements in metacognitive performance for our experimental condition.

Nonetheless, we still hypothesize that even under noisy conditions, lower impasse severities, higher self-monitoring accuracies, less bias, and better discrimination will be predictive of better cognitive performance. Thus we again computed partial correlations with posttest over all students, as originally reported in Forbes-Riley and Litman (2010). With the exception of discrimination, the ITSPOKE-AUTO correlations shown in the last two columns of Table 25.2 replicate the ITSPOKE-WOZ correlations of Table 25.1. Other comparisons between our wizarded and automated results (e.g., learning correlations with additional independent measures and regressions with multiple independent measures) can be found in Forbes-Riley and Litman (2010).

[10] In contrast, a measure of impasse *resolution* might positively correlate with learning, as resolving an impasse could reduce the severity of future impasse opportunities. In a prior ITSPOKE experiment, we in fact improved student learning by detecting and re-remediating one particular type of unresolved incorrectness impasse (Rotaru & Litman, 2009).

[11] The p-values for the 4 ANOVAs comparing the metacognitive metrics were respectively 0.83, 0.75, 0.72, 0.91. Due to both these extremely high p-values, and the fact that the means were not as predicted, we did not run the paired comparisons.

Discussion

We presented an analysis of student metacognitive performance using data from both wizarded and fully automated dialogue tutors that adapt to student uncertainty. The performance measures examined include several measures of metacognitive performance taken from various literatures but have been adapted for our tutorial dialogue context by computing them from tutor annotations of student uncertainty and correctness. We also introduce a new learning impasse severity measure derived from a theory of uncertainty and incorrectness as learning impasses. While in prior work we demonstrated that remediating after uncertainty impasses improves learning in both wizarded and fully automated conditions (Forbes-Riley & Litman 2011a, 2011b), our results here suggest that further investigation into better ways of remediating student uncertainty holds promise for further improving student cognitive as well as metacognitive performance.

With respect to improving cognitive performance, our correlation results suggest that if we can enhance our tutor to improve metacognitive performance, we may also further improve cognitive performance. Our correlations show that (tutor perception of) **impasse severity**, **self-monitoring accuracy**, and **bias** significantly or as a trend predict student learning (negatively, positively, and negatively, respectively) in both our wizarded and fully automated corpora. 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.

With respect to improving metacognition, our ANOVA results suggest that under upper-bound wizarded conditions, remediating student uncertainty holds promise for improving student metacognitive abilities (in our study, impasse severity and self-monitoring accuracy). However, the results with ITSPOKE-AUTO suggest that achieving this potential will require very high performing speech and language components.

In particular, while our ANOVAs for ITSPOKE-WOZ 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 *non-Adapt* (the original system). We also see similar results for *Simple* and *Random* compared to *non-Adapt* 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 & Roll 2007, Roll & Aleven 2008, Saadawi et al. 2009).

We found it surprising that neither experimental condition outperformed *Random*, even after we changed *Random* in ITSPOKE-AUTO to only adapt after CnonU answers (non-impasse states). Since a "nonuncertain" (nonU) answer may actually be certain or neutral, we hypothesize that adapting to CnonUs might still be effective at increasing certainty.

Finally, we recently found interactions between learning and user classes based on user domain expertise and gender in the wizarded corpus (Forbes-Riley & Litman, 2009b); we will investigate whether the interactions with these classes extend to the student metacognitive metrics discussed in this paper.

In conclusion, our work shows that the student speech signal holds important information about metacognition that most intelligent tutoring systems researchers have not yet mined. In particular, uncertainty is conveyed at least partially and sometimes most strongly through speech and tells us something about the student's accuracy of self-monitoring, which itself relates to learning. Although we have not yet attempted to dynamically adapt to metacognitive performance in our dialogue tutor to help students learn better at the cognitive level, or even improve metacognitive abilities, our results suggest that this is a plausible approach for future directions.

Acknowledgements This work is supported by NSF Award 0631930. A preliminary version of the wizarded results was presented at the AAAI Symposium on Cognitive and Metacognitive Educational Systems (Litman & Forbes-Riley, 2009a), while a subset of our automated results (the correlations) were previously published at ITS 2010 (Forbes-Riley & Litman, 2010).

References

Aist, G., Kort, B., Reilly, R., Mostow, J., & 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 *Proceedings of Intelligent Tutoring Systems Workshop on Empirical Methods for Tutorial Dialogue Systems* (pp. 483–490), San Sebastian, Spain.

Aleven, V., & Roll, I. (Eds.) (2007). *AIED workshop on metacognition and self-regulated learning in intelligent tutoring systems*.

Brennan, S.E., & 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,34*, 383–398.

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

Dijkstra, C., Krahmer, E. J., & Swerts, M. (2006). Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of Speech Prosody 2006*, Dresden: TUDpress.

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

Forbes-Riley, K., & Litman, D. (2009a). Adapting to student uncertainty improves tutoring dialogues. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* AIED 2009, July 6–10, 2009, Brighton, UK. Frontiers in Artificial Intelligence and Applications 200 IOS Press 2009, ISBN 978-1-60750-028-5.

Forbes-Riley, K., & Litman, D. (2009b). A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association*, Brighton, UK.

Forbes-Riley, K., & Litman, D. (2010). Metacognition and learning in spoken dialogue computer tutoring. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS)*, Pittsburgh, PA.

Forbes-Riley, K., & Litman, D. (2011a). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication, 53*(9–10), 1115–1136.

Forbes-Riley, K., & Litman, D. (2011b). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language, 25*(1), 105–126.

Forbes-Riley, K., Litman, D., & Rotaru, M. (2008a). 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.

Forbes-Riley, K., Rotaru, M., & Litman, D. (2008b). 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

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F.,& Rosenfeld, R. (1993). The SPHINX-II speech recognition system: An overview. *Computer Speech and Language,2*, 137–148.

Jordan, P., Hall, B., Ringenberg, M., Cui, Y., and Rosé, C. (2007). Tools for authoring a dialogue agent that participates in learning studies. In *Artificial Intelligence in Education (AIED)*, pp. 43–50.

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

Liscombe, J., Venditti, J., & Hirschberg, J. (2005, September). Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech/Eurospeech Conference on Speech Communication and Technology*, Lisbon, Portugal.

Litman, D., & Forbes-Riley, K. (2009a). Improving (meta) cognitive tutoring by detecting and responding to uncertainty. Technical Report FS-09-0:Cognitive and Metacognitive Educational Systems: Papers from the AAAI Symposium, AAAI Arlington, VA, November.

Litman, D., & Forbes-Riley, K. (2009b). Spoken tutorial dialogue and the feeling of another's knowing. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, London.

Litman, D., Moore, J., Dzikovska, M., & Farrow, E. (2009, July). Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* AIED 2009, Brighton, UK. Frontiers in Artificial Intelligence and Applications 200 IOS Press 2009, ISBN 978-1-60750-028-5.

Nicholas, G., Rotaru, M., & Litman, D. J. (2006). Exploiting word-level features for emotion prediction. In *Proceedings of IEEE/ACL Workshop on Spoken Language Technology*, Aruba.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement,66*, 258–271.

Pon-Barry, H. (2008). Prosodic manifestations of confidence and uncertainty in spoken language. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association,* September 2008 (pp. 74–77). Brisbane, Australia.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education,16*, 171–194.

Roll, I., & Aleven, V., (Eds.) (2008). *ITS Workshop on meta-cognition and self-regulated rearning in educational technologies*.

Rotaru, M., & Litman, D. J. (2009). Discourse structure and performance analysis: Beyond the correlation. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, London.

Saadawi, G. M. E., Azevedo, R., Castine, M., Payne, V., Medvedeva, O., Tseytlin, E., Legowski, E., azen Jukic, D., & Crowley, R. S. (2009). Factors affecting feeling-of-knowing in a medical intelligent tutoring system: The role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education,15*, 9–30.

Smith, V. L. and Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language,32*, 25–38.

Tsukahara, W., & Ward, N. (2001). Responding to subtle, fleeting changes in the user's internal state. In *Proceedings of SIG-CHI on Human Factors in Computing Systems* (pp.77–84).

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., & Wilson, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp.158–167).

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

Witten, I. H., & Frank, E. (1999). Data Mining: Practical Machine Learning Tools and Techniques, I. H. Witten & E. Frank and Mark Hall, January 2011, Morgan Kaufmann Publishers (ISBN: 978-0-12-374856-0).