Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 25 (2011) 105–126
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system

Kate Forbes-Riley [a,*], Diane Litman [b]

[a] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260, USA
[b] Learning Research and Development Center, Computer Science Dept., University of Pittsburgh, Pittsburgh, PA 15260, USA

Received 5 May 2008; accepted 15 December 2009
Available online 28 December 2009

Abstract

We describe the design and evaluation of two different dynamic student uncertainty adaptations in wizarded versions of a spoken dialogue tutoring system. The two adaptive systems adapt to each student turn based on its uncertainty, after an unseen human "wizard" performs speech recognition and natural language understanding and annotates the turn for uncertainty. The design of our two uncertainty adaptations is based on a hypothesis in the literature that uncertainty is an "opportunity to learn"; both adaptations use additional substantive content to respond to uncertain turns, but the two adaptations vary in the complexity of these responses. The evaluation of our two uncertainty adaptations represents one of the first controlled experiments to investigate whether substantive dynamic responses to student affect can significantly improve performance in computer tutors. To our knowledge we are the first study to show that dynamically responding to uncertainty can significantly improve learning during computer tutoring. We also highlight our ongoing evaluation of our uncertainty-adaptive systems with respect to other important performance metrics, and we discuss how our corpus can be used by the wider computer speech and language community as a linguistic resource supporting further research on effective affect-adaptive spoken dialogue systems in general.

Keywords: Affective states in spontaneous data; Spoken dialogue tutoring system; Automatic affect adaptation; Wizard of Oz experimental design; Collection and annotation of realistic, representative, and publicly

1. Introduction

As research on developing more natural and effective human–computer interaction progresses into more complex domains, there has been increasing interest in incorporating information about affective states into the interaction model. For example, within spoken dialogue systems research, a wide range of linguistic information, including pitch, energy, and timing information, has been successfully extracted from the user's speech signal and larger dialogue context and used to automatically detect user affective states (Batliner et al., 2008;

* Corresponding author. Tel.: +1 412 624 1261.
E-mail addresses: forbesk@cs.pitt.edu (K. Forbes-Riley), litman@cs.pitt.edu (D. Litman).

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2009.12.002

D'Mello et al., 2008; Litman and Forbes-Riley, 2006a; Vidrascu and Devillers, 2005; Lee and Narayanan, 2005; Shafran et al., 2003). Similarly, within the domain of tutoring systems, sophisticated models of learner affective states have been developed that take not only learner-based cues (e.g., linguistic, visual, and physiological), but also the learning environment into account (D'Mello et al., 2008; Porayska-Pomsta et al., 2008; Conati and Maclaren, 2004; Gratch and Marsella, 2003; de Vicente and Pain, 2002). The larger goal of such automatic affect detection is to enable automatic affect adaptation; these spoken dialogue and tutoring system researchers hypothesize that responding to user affect will significantly improve system performance. However, to date most deployed spoken dialogue systems and computer tutors ignore user affective states when determining how to respond. Moreover, it is still largely an open question as to what are the most effective methods of adapting to user affective states. To some extent the answer to this question depends on the task domain, the system performance metric being targeted for improvement, and the affective state(s) being targeted for adaptation.

Within the tutoring spoken dialogue system domain, student learning is the primary system performance metric, and student uncertainty is an affective state of primary interest. Although uncertainty is not one of the "big 6" basic emotions such as anger and happiness (Ekman and Friesen, 1978), tutoring dialogue research suggests uncertainty plays an important role in the learning process; in particular, it has been related to correctness and learning (Craig et al., 2004; Bhatt et al., 2004). Due to the complexity of the information exchange in a tutoring dialogue, *dynamic* affect-adaptive tutoring systems-that is, systems that recognize and respond to user affect on a turn by turn basis-are often modeled on human tutors' responses. In addition, most affect-adaptive computer tutors have been evaluated within a "Wizard of Oz" scenario, where a human "wizard" performs system tasks such as speech recognition, natural language understanding, and affect detection. Wizarding the system in this way removes noise that might have potentially distracted from the dialogue interaction due to misrecognition of the user's utterance and/or affective state; thus the system design is tested under the best possible conditions. For example, in Tsukahara and Ward (2001), positive feedback responses to various affective states in student answers, including praising acknowledgments after uncertain answers, were developed based on a frequency analysis of human tutor responses. These responses were implemented in a Memory Game computer tutor, in which a human wizard performed speech recognition and natural language understanding (affect detection was automatic). Students rated the usability of the resulting affect-adaptive system more highly than a non-adaptive version, but no significant differences in student learning were reported. Similarly, in Aist et al. (2002), a human wizard performed speech recognition and natural language understanding in a spoken Reading Tutor, and then provided "emotional scaffolding" (e.g., "Good try") after detecting various affective states in student answers, including uncertainty. The emotional scaffolding resulted in increased student persistence, but did not yield improved learning. As these examples illustrate, dynamically adapting to student affect with positive feedback and/or empathy has yielded improvements for performance metrics such as user satisfaction, but has not yet yielded significant learning improvements. Other dialogue system domains have also shown performance improvements by dynamically adapting to user affect with positive feedback and/or empathy (Liu and Picard, 2005; Klein et al., 2002; Prendinger and Ishizuka, 2001). For example, Liu and Picard (2005)'s health assessment system responds with empathy to instances of user stress. Similarly, Klein et al. (2002)'s gaming system responds with sympathy and apology to instances of user frustration. In both of these studies, user satisfaction is considered a primary metric of system performance, and both studies successfully showed that users preferred to use the adaptive system over non-adaptive versions.

There have also been a number of non-dynamic (*static*) approaches to affect adaptation in tutoring systems. These systems employ "empathetic agents", whose responses take student affect into account but do not detect and respond to the specific affective state of each student turn (Wang et al., 2008, 2005; Hall et al., 2004); in some cases significant learning improvements have been achieved. In particular, Wang et al. (2008, 2005) implement a model of "socially intelligent tutoring" based on politeness theory in an online learning system. They conduct a series of Wizard of Oz studies in which students either used the socially intelligent system and received polite tutorial feedback after every turn, or used the control system and received direct feedback after every turn. The socially intelligent system was found to yield increased student learning as compared to the control system. Related to this work is recent research on natural language generation in dialogue systems that addresses automatic generation of different system personality styles, such as in Mairesse and Walker (2008).

Taken together, this prior research suggests that while computer tutors should be made more polite and empathetic overall to increase their learning effectiveness, dynamically responding to student uncertainty with positive feedback and empathy is not sufficient in isolation to increase student learning. Tutoring theories addressing the relationship between uncertainty and learning (see Section 3) suggest that more substantive dynamic system responses to uncertainty over and above correctness might be more effective at increasing learning. To our knowledge, only one other study has evaluated substantive dynamic computer tutor adaptations to affective student turns. In particular, Pon-Barry et al. (2006) use a frequency analysis to extract two human tutor responses to uncertain answers (correct and incorrect, respectively) from a human-tutoring corpus. These responses ("paraphrasing" after correct + uncertain answers and "referring back to past dialogue" after incorrect + uncertain answers) were then implemented and evaluated in the fully-automated SCoT-DC spoken dialogue tutor. When used after all correct and incorrect answers, the adaptations were found to significantly increase learning as compared to not using them at all. However, when used *only* after correct + uncertain and incorrect + uncertain answers, the adaptations did not significantly increase learning as compared to not using them at all. This result suggests that the adaptations did not actually target uncertainty; rather, they improved the overall tutoring strategy for responding to (in)correctness. This might be due to the fact that the adaptations were not based on statistically significant differences in how a human tutor responds to uncertain versus non-uncertain answers. In addition, because the tutoring system was fully automated rather than wizarded, speech recognition and understanding errors may have negatively impacted the effectiveness of the adaptations. Moreover, the system used only a limited set of features to recognize uncertainty, which may also have decreased the effectiveness of the adaptations.

Although there are to date so few controlled evaluations of substantive dynamic adaptations to user affect in the spoken dialogue tutor domain, similar evaluations performed in other spoken dialogue system domains have shown that substantive dynamic adaptations can significantly improve system performance; however, the adaptations being evaluated were based on recognition of communication problems rather than recognition of user affect. For example, in Litman and Pan (2002), dialogue strategies were automatically adapted dynamically based on repeated speech recognition errors. The adaptive system outperformed the non-adaptive system for novice users by significantly increasing the task completion rate. In Chu-Carroll and Nickerson (2000), initiative strategies were automatically adapted dynamically based on participant roles, features of the current utterance such as ambiguity, and dialogue history. The adaptive system outperformed the non-adaptive system in terms of system usability, dialogue efficiency, and dialogue quality measures. Reinforcement learning is another technique that has produced substantive dynamic adaptations based on user states in spoken dialogue systems (see Section 8).

In this paper we discuss the design of two different substantive dynamic adaptations to student uncertainty in our spoken dialogue computer tutor, and the evaluation of these adaptations via controlled experiment using a Wizard of Oz scenario, where the wizard performed speech recognition, natural language understanding, and uncertainty recognition. Both uncertainty adaptations were derived from tutoring theory that views both uncertainty and incorrectness as a "learning opportunity": the *Simple* adaptation provided additional tutoring content after every uncertain or incorrect student answer, to take advantage of all learning opportunities. The *Complex* adaptation provided the same additional tutoring content after all learning opportunities, but varied the presentation of this content based on a statistical analysis of human tutor dialogue act responses to uncertainty and incorrectness, and it also provided empathetic feedback after every uncertain or incorrect answer. In contrast, the original non-adaptive system ignored uncertainty-it only provided additional tutoring content after incorrect answers, and it only provided feedback acknowledging the answer's (in)correctness.

Our results show that our *Simple* adaptation significantly improved student learning as compared to a non-adaptive version of our wizarded computer tutor. To our knowledge we are the first study to show that dynamically responding to student uncertainty can significantly improve learning during computer tutoring. We also highlight our ongoing evaluation of our uncertainty-adaptive systems with respect to other important performance metrics that are important for spoken dialogue systems in general, including user satisfaction and dialogue-based metrics. Finally, we discuss how the corpus resulting from our experiment can be used by the wider computer speech and language community as a linguistic resource to support further research on developing effective affect-adaptive spoken dialogue systems in all domains.

This article is organized as follows: Section 2 describes our original non-adaptive spoken dialogue tutoring system and the Wizard of Oz version of our system. Section 3 discusses why student uncertainty is an important affective state to adapt to in such systems. Section 4 discusses how our two uncertainty adaptations are derived from tutoring theory, and explains how these adaptations are implemented in our system. Section 5 describes the controlled experiment and corpus collection using wizarded adaptive and non-adaptive versions of our system. Section 6 discusses our evaluation of the impact of the uncertainty adaptations on student learning. Section 7 details the resulting corpus, including speech files, transcriptions, annotations, and descriptive statistics, and also discusses other uses of the corpus, including our evaluations of the adaptive systems with respect to other performance metrics, and the use of our corpus by the wider community as a data mining resource for improving automatic affect detection and adaptation in spoken dialogue systems. Section 8 discusses our future work, including a future evaluation of a fully automated version of the system used in this experiment.

2. ITSPOKE: original and wizarded system versions

Our original (non-adaptive) tutoring system is called ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman and Forbes-Riley, 2006a). ITSPOKE is a fully automated spoken dialogue tutoring system that is built on top of the Why2-Atlas text-based tutoring system (VanLehn et al., 2002).

ITSPOKE tutors students in five qualitative physics problems. One physics problem is shown in Fig. 1. Each physics problem is tutored via a spoken dialogue interaction between student and ITSPOKE. Each dialogue consists of a series of questions about the topics needed to solve the physics problem. The dialogues have a Question–Answer–Response format, which is implemented with a finite state dialogue manager. In our ori-

Suppose a lightweight car and a massive truck are driving towards each other in a straight line. They hit a patch of frictionless ice and have a head-on collision. Upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? (As usual, assume air resistance is negligible).

Fig. 1. Example ITSPOKE physics problem.

TUTOR1: Upon which vehicle is the impact force greater?
STUDENT: the car *[Incorrect]*
TUTOR2: That doesn't sound right. *[FEEDBACK]* We just discussed that by Newton's Third law, when two objects collide, the forces they exert on each other are equal in magnitude and opposite in direction. This is true regardless of the objects' differing masses. *[BOTTOMOUT]*

Fig. 2. Example of a **BottomOut** response to an incorrect answer in the original (non-adaptive) ITSPOKE.

TUTOR1: What's the overall net force on the truck equal to?
STUDENT: Zero?? *[Incorrect]*
TUTOR2: That doesn't sound right. *[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? *[SUBDIALOGUE]*

Fig. 3. Example of a **Subdialogue** response to an incorrect answer in the original (non-adaptive) ITSPOKE.

ginal (non-adaptive) ITSPOKE, system responses (states) depend only on the correctness of the student answers (transitions between states).

If the student answers a question correctly, ITSPOKE responds with correctness feedback and then moves on to the next question. Correctness feedback is selected from the following phrases to indicate that the answer was correct: *Fine; That's correct; Excellent; That's right; Good; Right; Correct*.

If the student answers a question incorrectly, ITSPOKE responds with a remediation. This response begins with correctness feedback selected from the following phrases to indicate that the answer was incorrect: *That doesn't sound right; I don't think so; Well...; That's not what I expected*. The rest of the response takes one of two forms:

- For incorrect answers to questions about easier topics, ITSPOKE gives a **BottomOut**, i.e., provides the correct answer with a brief statement of reasoning. This is illustrated in Fig. 2.
- For incorrect answers to questions about harder topics, ITSPOKE engages the student in a **Subdialogue**, i.e., one or more additional questions that walk the student through the more complex line of reasoning required to achieve the correct answer. This is illustrated in Fig. 3 (only the first question in the subdialogue is shown).

2.1. ITSPOKE-WOZ: Wizard of Oz version of ITSPOKE

For this study, we used a Wizard of Oz version of ITSPOKE (ITSPOKE-WOZ). In ITSPOKE-WOZ, a few system components were replaced by a human "wizard": The wizard performed speech recognition, natural language understanding, and uncertainty annotation, for each student answer. In this way, we tested the upper bound performance of our uncertainty adaptations without any potentially negative impact of automated versions of these tasks.

Fig. 4 shows a screenshot of the wizard's interface used during the experiment. The top Problem Statement box shows the physics problem. The middle Dialogue History box shows a history of the text of the tutor turns. The student turns are not shown in this box because they are not transcribed until after the experiment.

Note that students use almost the same interface as the wizard, except that students only see the Problem Statement and Dialogue History boxes. Although students listen to the tutor speech through headphones (see Section 5), the Dialogue History gives them the option of reading along.

The lowest section of the interface is seen only by the wizard; it is used to annotate the student turns for correctness and uncertainty. Upon hearing each student answer, the wizard annotates whether or not it is uncertain in the Uncertain checkbox in the lower right. All annotated values are logged and sent to the dialogue manager to determine the system's response when the wizard clicks the "OK" box.

In the lower left selection area, the wizard annotates whether the heard answer is correct or incorrect, and to which category of answer it belongs: correct, unanticipated, and don't know (labeled $=c$, $=u$, and $=d$, respectively). For example, in Fig. 4, "same" is the correct answer category for the tutor question recorded in the Dialogue History pane. A variety of actual heard student answers would fall into this category, including "they are the same", "they are equal", "there's no difference between them", etc. The "unanticipated" category is used for incorrect answers. For example, in Fig. 4, a wide variety of student answers, including "they are different", "they are zero", and "the keys are falling faster than the man" would all fall into this category. The "don't know" category is used for answers such as "I have no idea", "um...", "I give up". Note that "don't know" answers are treated as incorrect in terms of system response content in the original ITSPOKE system; however, they receive special treatment in the *Complex* uncertainty adaptation (see Section 4.2).

3. Targeting student uncertainty

Our investigation into the development of dynamic substantive student affect adaptations initially targets a single student affective state: **uncertainty**. Although uncertainty does not fall within the "big 6" set of basic emotions described in Ekman and Friesen (1978) (fear, happiness, sadness, disgust, anger, surprise), tutoring researchers have argued that this set needs to be supplemented or even replaced to describe the range of emotions relevant to the learning process (D'Mello et al., 2008). In line with this literature, we use the term "affect"

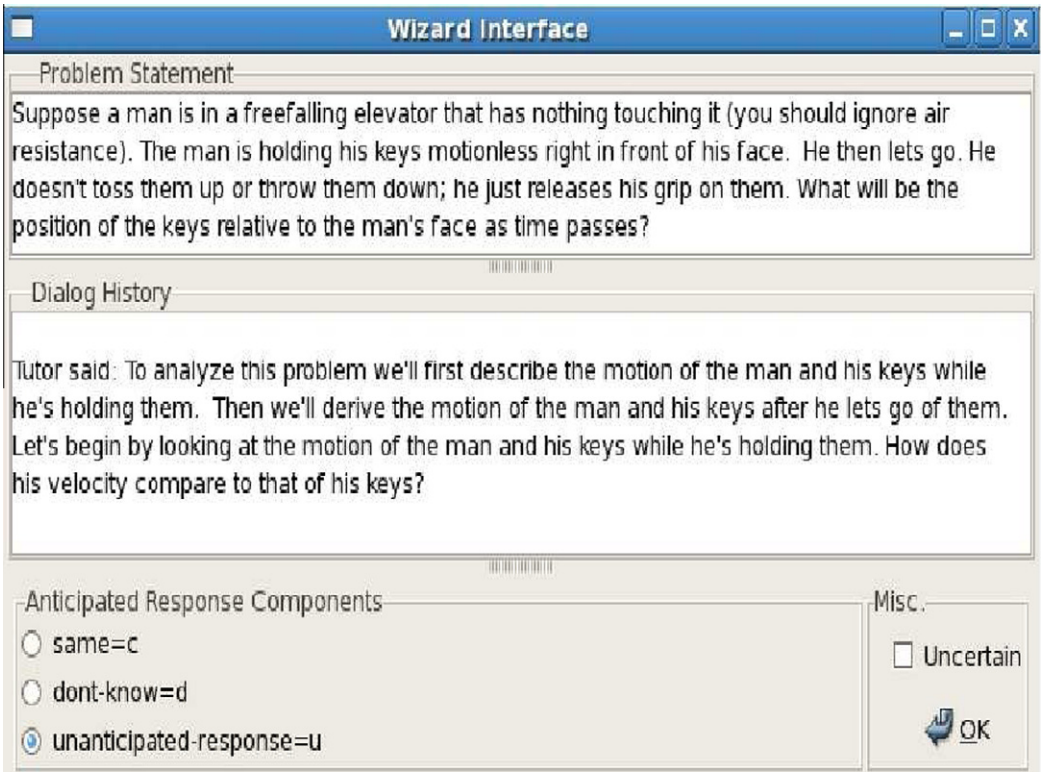

Fig. 4. Screenshot of wizard interface.

in this paper to cover both emotions and attitudes that may impact student learning and can affect how students communicate their answers during tutoring. Other tutoring researchers have also taken this complex view when annotating various non-traditional affective states in tutoring dialogues, including uncertainty or confusion (D'Mello et al., 2008; Pon-Barry et al., 2006; Bhatt et al., 2004) and confidence, self-efficacy, or flow (McQuiggan et al., 2008; D'Mello et al., 2008). Similarly, some speech researchers have found that the narrow sense of "emotion" is too restrictive because it excludes states in speech where affective state is present but not full-blown, including arousal and attitude (see Cowie and Cornelius, 2003).

We target uncertainty for two reasons. First, uncertainty occurred significantly more frequently than other student affective states in our previously collected ITSPOKE corpora (Litman and Forbes-Riley, 2004). Second, uncertainty is an affective state of primary interest in the tutoring dialogue domain because research suggests it plays an important role in the learning process. In particular, tutoring researchers hypothesize that uncertainty can signal to the tutor that there is an opportunity for constructive learning to occur, and that experiencing uncertainty can motivate a student to engage in learning (e.g., VanLehn et al., 2003; Kort et al., 2001). However, studies have also shown that uncertainty and incorrectness cannot be equated (Bhatt et al., 2004). In addition, studies have shown correlations between learning and confusion, with confusion being defined as an indicator of uncertainty. For example, Craig et al. (2004) use Rozin and Cohen (2003)'s definition of student confusion as indicating an uncertainty about what to do next or how to act, or a need for clarification or more information. Craig et al. (2004) argue that confusion therefore accompanies "cognitive disequilibrium" (Graesser and Olde, 2003), in which learners confront obstacles to goals, salient contrasts, equivalent alternatives, or other experiences that fail to match their expectations. The cognitive disequilibrium, and the confusion that accompanies it, has a high likelihood of causing deliberation and inquiry aimed at restoring cognitive equilibrium. Craig et al. (2004) report that the proportion of student confusion turns positively correlated with learning in the AutoTutor system (Graesser et al., 2005).

Nominal:	[Incor+Cert]	[Incor+Uncert]	[Cor+Uncert]	[Cor+Cert]
Scalar:	3	2	1	0
Severity:	most	less	least	none

Fig. 5. Different learning impasse state severities.

Tutoring dialogue research has also shown that student uncertainty can be annotated with a reasonable degree of interannotator reliability. For example, Bhatt et al. (2004) annotated student turns containing lexical expressions of "hedging" (e.g., "I guess", "um") with an inter-annotator reliability of 0.97 Kappa.[1] In our prior ITSPOKE corpora, two annotators labeled 4533 student turns for binary uncertainty (uncertain versus certain), yielding an inter-annotator reliability of 92% (0.73 Kappa).[2]

Our annotation scheme for labeling student uncertainty distinguishes two labels. The *Uncertain* label is used for answers expressing uncertainty or confusion about the material being tutored (e.g., as opposed to students' uncertainty directed towards themselves or towards being tutored by a computer). The *Certain* label is used for all other answers (which actually may be either certain or neutral for certainty).

Our uncertainty annotation is based wholly on the annotator's subjective judgment. S/he can base this decision on evidence from many knowledge sources, including lexical items (e.g., "I don't know"), disfluencies (e.g.,"um", "gra-gravity"), sentence fragments (e.g., "gravity is. . ."), and acoustic-prosodic features (e.g., intonation, tempo, energy) in the student speech, as well as the larger dialogue context (e.g., the student's past performance and past frequency of expressed uncertainty). However, such evidence is used on a speaker-dependent basis because particular cues are not used consistently or unambiguously across speakers.

Note finally that the best way to manually label affective states is still an open question. For example, as we have done, many researchers rely on independent judges (e.g., Porayska-Pomsta et al., 2008; Shafran et al., 2003; Ang et al., 2002; Narayanan, 2002), while others use self-reports (e.g., McQuiggan et al., 2008; D'Mello et al., 2008; Yannakakis et al., 2008; Klein et al., 2002). D'Mello et al. (2008) compare multiple types of labeling, including self-reports, peer labelers, and trained judges, as well as consensus labelings derived from various pairings of labelers. Batliner et al. (2008) use majority voting (similar to consensus labeling); they also provide a detailed discussion of the complications underlying affective state annotation.

4. Developing dynamic substantive uncertainty adaptations for ITSPOKE-WOZ

We derived two dynamic substantive system adaptations for student uncertainty from a hypothesis in the tutoring dialogue literature that uncertainty and incorrectness both signal learning opportunities, called "learning impasses": opportunities for the student to better learn the material about which s/he is uncertain or incorrect (VanLehn et al., 2003). VanLehn et al. (2003) argue that the learning impasse motivates the student to take an active role in constructing a better understanding of the material being tutored.

However, we observed that in order to be motivated to resolve a learning impasse, the student must first perceive that the impasse exists. Incorrectness and uncertainty differ in terms of this perception. Incorrectness simply signals that the student has reached a learning impasse, while uncertainty-in both a correct and a incorrect answer-signals that the student perceives that they have reached a learning impasse. Based on this distinction, we associated each of the four possible student answer combinations of binary uncertainty (**Uncert**, **Cert**) and binary correctness (**Incor**, **Cor**) with a scalar value from 3 to 0, as shown in Fig. 5. We hypothesized that these scalar values correspond to the severity of the student's learning impasse state with respect to his/her current answer (Forbes-Riley et al., 2008a). In particular, "0" corresponds to a state in which the student

[1] Although interpreting Kappa values is somewhat controversial and varies depending on the application field, we find the following agreement standard (Landis and Koch, 1977) to be a useful guideline: 0.21–0.40 = "Fair"; 0.41–0.60 = "Moderate"; 0.61–0.80 = "Substantial"; 0.81–1.00 = "Almost Perfect".

[2] See Forbes-Riley et al. (2008c) for further discussion of this annotation. Other studies of affective state in naturally occurring dialogues in other domains have yielded Fair to Substantial Kappas for other affective state labels (e.g., Ang et al., 2002; Narayanan, 2002; Shafran et al., 2003).

is not experiencing an impasse, because s/he is correct and not uncertain about the answer. "3" corresponds to a state in which the student is experiencing the most severe type of impasse, because s/he is incorrect and is not aware of it. "2" and "1" correspond to states of lesser severity: the student is incorrect but aware that s/he might be incorrect, and the student is correct but uncertain about whether s/he is correct, respectively. In Forbes-Riley et al. (2008a) we show empirical support for distinguishing impasse severities; we show that total and average learning impasse severity are both significantly negatively correlated with student learning.

As discussed in Section 2, both types of incorrectness impasse ([Incor + Cert] and [Incor + Uncert]) already receive additional substantive content via a BottomOut or Subdialogue in the non-adaptive ITSPOKE-WOZ. However, one type of uncertainty impasse is ignored: [Cor + Uncert].

The hypothesis underlying our two uncertainty adaptations is thus that student learning should increase if the adaptive versions of our ITSPOKE-WOZ provide additional substantive content to remediate *every* learning impasse: [Cor + Uncert], [Incor + Uncert], and [Incor + Cert].

4.1. Simple uncertainty adaptation

Our *Simple* uncertainty adaptation represents the simplest instantiation of this hypothesis: give all learning impasses *the same additional substantive content*. Therefore, our *Simple-adaptive* ITSPOKE-WOZ responds to [Cor + Uncert] answers in the same way as [Incor + Uncert] and [Incor + Cert] answers: for any given question, all three answer types receive the same BottomOut or Subdialogue response. This is illustrated in Fig. 6. As shown, the [Cor + Uncert] answer in **STUDENT** receives a Subdialogue response in **TUTOR2**. Comparison with Fig. 3 shows that this is the same Subdialogue response that incorrect answers already receive in non-adaptive ITSPOKE-WOZ.

Implementing Simple-adaptive ITSPOKE-WOZ involved changing the next state transitions in ITSPOKE-WOZ's finite state dialogue manager. Instead of transitioning based only on the correctness value of the student answer, the transition is based on the answer's combined correctness and uncertainty value: [Cor + Uncert], [Incor + Uncert] and [Incor + Cert] answers all transition to the same Subdialogue or BottomOut response.

Note that Simple-adaptive ITSPOKE-WOZ does not change the feedback given after any student answer. Identically to non-adaptive ITSPOKE-WOZ, correct answers still receive feedback selected from the following phrases: *Fine; That's correct; Excellent; That's right; Good; Right; Correct*, and incorrect answers still receive feedback selected from the following phrases: *That doesn't sound right, I don't think so, Well. . ., That's not what I expected.*

Thus, Simple-adaptive ITSPOKE-WOZ differs from non-adaptive ITSPOKE-WOZ only in terms of its substantive response to [Cor + Uncert] answers. Simple-adaptive ITSPOKE-WOZ and non-adaptive ITSPOKE-WOZ are identical in terms of their response to [Incor + Uncert], [Incor + Cert] and [Cor + Cert] answers.

4.2. Complex uncertainty adaptation

Our *Complex* uncertainty adaptation revises our Simple adaptation based on statistical analysis of human tutor responses to student uncertainty and correctness. In particular, all learning impasses are again given *the same additional substantive content*; however, *the dialogue act* used to present this content and *the feedback*

TUTOR1: What's the overall net force on the truck equal to?
STUDENT: The force of the car hitting it?? *[Cor+Uncert]*
TUTOR2: 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? *[SUBDIALOGUE]*

Fig. 6. Example of **Simple-adaptive** ITSPOKE-WOZ's adaptation for **[Cor + Uncert]** answers.

Table 1
Examples of variations of feedback phrases used in complex uncertainty adaptations for different student answer impasse severities.

Impasse type	Examples of feedback phrases used
[Cor + Cert]	That's right
[Cor + Uncert]	That's right, but you don't sound very certain, so let's recap
[Incor + Uncert]	Good try, but that's not right. It sounds like you knew there might be an error in your answer. Let's fix it
[Incor + Cert]	I'm sorry, but there's a mistake in your answer that we need to work out

used to acknowledge the answer both vary depending on the answer's learning impasse severity (as defined above in Fig. 5).

4.2.1. Feedback variations in the complex adaptation

Our feedback variations were based on multiple prior computer tutor results discussed in Section 1, which showed that empathetic system responses can positively impact student performance. Our new feedback phrases acknowledged *both* the propositional content (correctness) and the affective content (uncertainty) in student answers. We authored at least five new feedback phrases for each learning impasse type. Table 1 provides one example for each impasse type; additional variations are shown below in Figs. 7–10.

As exemplified in Table 1, our feedback phrases for [Cor + Cert] are selected from the same feedback list used for correct answers in non-adaptive ITSPOKE-WOZ. We decided not to explicitly acknowledge the lack of a learning impasse in [Cor + Cert] answers because a pilot study showed that this feedback quickly became annoying to students (e.g., "That's correct and you don't sound uncertain so let's move on.")

Our feedback phrases for [Cor + Uncert] and [Incor + Uncert] assert that the answer is correct or assert that it is incorrect in an empathetic manner, while also acknowledging that uncertainty was detected. In this way the system explains why it will be providing additional substantive content for this turn.[3]

Our feedback phrases for [Incor + Cert] assert that the answer is incorrect in an empathetic manner, in order to help the student recognize that a learning impasse has been reached and motivate him/her to resolve it.

4.2.2. Dialogue act variations in the complex adaptation

Our dialogue act variations of the substantive response content for the different learning impasse types were based on a statistical analysis of human tutor dialogue behavior. In particular, we used the χ^2 test to find dependencies between different learning impasse severities of student answers and nine dialogue act responses that a human tutor used in the turn immediately following these answers. Our dialogue acts are based on similar schemes from other tutorial dialogue projects (e.g., Graesser et al., 1995). This analysis is summarized below, and discussed in greater detail in Forbes-Riley and Litman (2007).

Our χ^2 analysis yielded a few impasse–dialogue act response pairs that occurred significantly more or less than expected by chance. Although the χ^2 test is not a causal test; we formulated hypotheses about the reasons underlying these significant dependencies, and from these hypotheses derived specific system adaptations for each impasse type.

Our first set of dependencies are shown in Table 2. The first column shows the impasse–dialogue act response pair, and the second and third columns show the observed and expected values for this pair. Comparison of the observed and expected values gives the direction of the dependency, which is shown in the last column: a "−" indicates a dependency where the observed count is significantly less than expected; a "+" indicates a dependency where the observed count is significantly more than expected.

As shown in Table 2, the human tutor responded with a Short Question significantly less than expected after [Cor + Uncert] answers and significantly more after [Cor + Cert] answers ($p \leqslant .05$). The uncertainty

[3] Note that "don't know" answers (discussed in Section 2.1) are treated as [Incor + Uncert], because if a student asserts that s/he does not know the answer s/he has already perceived a learning impasse. However, in *Complex-adaptive* ITSPOKE-WOZ "don't know" answers receive specific feedback phrases, such as "I'm sorry you don't know, but don't give up. Let's think about it further".

TUTOR1: What's the overall net force on the truck equal to?
STUDENT: The impact force of the car on the truck?? *[Cor+Uncert]*
TUTOR2: 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 that 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 LONG BOTTOM OUT]*

Fig. 7. Example of **Complex-adaptive** ITSPOKE-WOZ's BottomOut adaptation for **[Cor + Uncert]** answers.

TUTOR1: What's the overall net force on the truck equal to?
STUDENT: Zero?? [Incor+Uncert]
TUTOR2: 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 SHORT 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? *[EXISTING SUBDIALOGUE]*

Fig. 8. Example of **Complex-adaptive** ITSPOKE-WOZ's BottomOut + Subdialogue adaptation for **[Incor + Uncert]** answers.

TUTOR1: Upon which vehicle is the impact force greater?
STUDENT: the car *[Incor+Cert]*
TUTOR2: I'm sorry, but I see an error in your answer. Let's fix it. *[FEEDBACK]* The problem statement says the truck and car collide. Newton's Third law says that the car and the truck both exert an impact force on each other. So according to Newton's Third law, how do the magnitudes of these forces compare? *[NEW SHORT SUBDIALOGUE] (followed by existing BottomOut)*

Fig. 9. Example of **Complex-adaptive** ITSPOKE-WOZ's new subdialogue adaptation for **[Incor + Cert]** answers.

in an [Cor + Uncert] answer indicates that the student has erroneously perceived that s/he might be incorrect; we hypothesized that the human tutor helped the student resolve this learning impasse before asking another question. In contrast, the tutor moved on to the next question after a [Cor + Cert] answer because no learning impasse needs to be resolved.

Our interpretation of the [Cor + Uncert]–Short Question dependency suggests that Complex-adaptive ITSPOKE-WOZ should not use a question to respond to [Cor + Uncert] answers. Recall that in non-adaptive ITSPOKE-WOZ, all substantive responses take one of two dialogue act forms (Section 2): BottomOut, which consists of one or more statements and is used after answers to easier questions, and Subdialogue, which consists of one or more questions and is used after answers to harder questions. Therefore we decided Complex-adaptive ITSPOKE-WOZ should use a BottomOut to respond to all [Cor + Uncert] answers. We reused the existing BottomOut responses from non-adaptive ITSPOKE-WOZ for [Cor + Uncert] answers to easier questions. Since only a Subdialogue response already existed for [Cor + Uncert] answers to harder questions, we authored a new BottomOut version of this Subdialogue. These new BottomOuts were typically long, because multiple statements were needed to summarize the content of a Subdialogue.

As an example, Fig. 7 shows how the Subdialogue used by non-adaptive ITSPOKE-WOZ in Fig. 3 has been changed to a long BottomOut for use in Complex-adaptive ITSPOKE-WOZ to respond to [Cor + Uncert]

TUTOR1: What will the velocity of the object be a second after that (where the initial velocity is 9.8m/s and the acceleration is 9.8m/s^2)?
STUDENT1: Is it 19.6 m/s? *[Cor+Uncert]*

NonAdapt Condition

TUTOR2: Good. Now back to the man and his keys. We've shown that they both have an acceleration of 9.8m/s^2. So, how do their velocities compare with each other, at every second during the fall?

Simple Condition

TUTOR2: Good. A second later, its velocity will be 9.8m/s + 9.8m/s = 19.6m/s. This is because its acceleration tells us that every second, its velocity increases by 9.8m/s. So, what'll its velocity be, a second after that?
STUDENT2: Um. I'm thinking. 29.4 m/s. *[Cor+Cert]*
TUTOR3: Right. Now back to the man and his keys. We've shown that they both have an acceleration of 9.8m/s^2. So, how do their velocities compare with each other, at every second during the fall?

Complex Condition

TUTOR2: Very good. But you seem hesitant, so let's recap. A second later, its velocity will be 9.8m/s + 9.8m/s = 19.6m/s. This is because its acceleration tells us that every second, its velocity increases by 9.8m/s. Now back to the man and his keys. We've shown that they both have an acceleration of 9.8m/s^2. So, how do their velocities compare with each other, at every second during the fall?

Random Condition

TUTOR2: *[Answer Randomly Selected = NO]* Good. Now back to the man and his keys. We've shown that they both have an acceleration of 9.8m/s^2. So, how do their velocities compare with each other, at every second during the fall?

Fig. 10. Corpus excerpts illustrating differing system responses across conditions.

Table 2
Student answer–Short Question dependencies ($p \leqslant .05$).

Impasse–dialogue act response pair	Observed	Expected	Direction
[Cor + Uncert]–Short Question	104	121	−
[Cor + Cert]–Short Question	285	268	+

Table 3
Student answer–BottomOut dependencies ($p \leqslant .05$).

Impasse–dialogue act response pair	Observed	Expected	Direction
[Incor + Uncert]–BottomOut	82	72	+
[Incor + Cert]–BottomOut	57	67	−

answers. Table 3 shows the second set of dependencies resulting from our χ^2 analysis. As shown, the human tutor responds with a BottomOut significantly more than expected after [Incor + Uncert] answers and significantly less than expected after [Incor + Cert] answers ($p \leqslant .05$). The uncertainty in an [Incor + Uncert] answer indicates that the student has already perceived that s/he might be incorrect; we hypothesized that the human tutor therefore immediately helped resolve the learning impasse with a BottomOut. In contrast, the lack of uncertainty in an [Incor + Cert] answer indicates that the student has not already perceived that s/he might be incorrect; the human tutor therefore helped the student perceive this learning impasse before

supplying the correct answer. Our interpretation of the [Incor + Uncert]–BottomOut dependency suggests that Complex-adaptive ITSPOKE-WOZ should use a BottomOut to respond to all [Incor + Uncert] answers. For [Incor + Uncert] answers to easier questions we reused the existing BottomOut responses from non-adaptive ITSPOKE-WOZ. For [Incor + Uncert] answers to harder questions, only a Subdialogue existed in non-adaptive ITSPOKE-WOZ; in these cases we used *both* a new short BottomOut version of the Subdialogue *and* the Subdialogue. Our reasoning here was that for [Incor + Uncert] answers to harder questions, a BottomOut may not suffice to fully resolve both the uncertainty and the incorrectness; instead a new short BottomOut first simply showed the final solution, then the Subdialogue walked the student through the steps to this solution. Thus our new BottomOuts for [Incor + Uncert] answers to harder questions were not the same as those we authored for [Cor + Uncert] answers.

As an example, Fig. 8 shows how the Subdialogue used by non-adaptive ITSPOKE-WOZ in Fig. 3 has been modified in Complex-adaptive ITSPOKE-WOZ for [Incor + Uncert] answers. A new short BottomOut gives the correct answer, then the Subdialogue walks the student through the reasoning.

Our interpretation of the [Incor + Cert]–BottomOut dependency suggests that Complex-adaptive ITSPOKE-WOZ should not use a BottomOut to respond to [Incor + Cert] answers. We thus decided to use a Subdialogue to respond to all [Incor + Cert] answers. For [Incor + Cert] answers to harder questions we reused the existing Subdialogues from non-adaptive ITSPOKE-WOZ. Our reasoning here was that the existing Subdialogue would help students first perceive and then resolve the learning impasse by walking them through the complex line of reasoning without immediately giving away the correct answer. For [Incor + Cert] answers to easier questions where only a BottomOut existed in non-adaptive ITSPOKE-WOZ, we used *both* a new short Subdialogue version of the BottomOut *and* the BottomOut. Our reasoning here was that the new Subdialogue, which consisted of a single easy question, would help students first perceive the impasse and give them a chance to supply the correct answer, which would then be reinforced and explained in the BottomOut.

As an example, Fig. 9 shows how the BottomOut used by non-adaptive ITSPOKE-WOZ in Fig. 2 has been modified in Complex-adaptive ITSPOKE-WOZ for [Incor + Cert] answers. A new Subdialogue reasks an easier version of the question in **TUTOR1** after reminding the student of the important concepts to consider. After the student answers this question (correctly or incorrectly), the existing BottomOut shown in Fig. 2 is given.

Table 4 summarizes how student answers with different learning impasse severities receive different dialogue act formats of the same substantive response across the non-adaptive, *Simple-adaptive*, and *Complex-adaptive* ITSPOKE-WOZs.

5. The experiment

To investigate the impact of the *Simple-adaptive* and *Complex-adaptive* systems on student performance, we performed a controlled experiment comparing the two uncertainty-adaptive systems with two control systems. The experiment had four conditions: two control conditions in which uncertainty was ignored by the system and two experimental conditions in which uncertainty was dynamically adapted to by the system. Note that uncertainty was manually labeled by the wizard and logged in all four conditions.

In the **Non-adaptive control condition** (*NonAdapt*), student used *Non-adaptive* ITSPOKE-WOZ, which was discussed in Section 2.

Table 4
Summary of variations in dialogue act format of substantive response content across non-adaptive, *Simple-adaptive*, and *Complex-adaptive* ITSPOKE-WOZs for different student answer impasse severities.

Impasse severity	Dialogue act format of existing substantive response		
	Non-adapt	*Simple*	*Complex*
[Cor + Cert]	NONE	NONE	NONE
[Cor + Uncert]	NONE	Existing format	Existing format if BottomOut, else new long BottomOut
[Incor + Uncert]	Existing format	Existing format	Existing format if BottomOut, else new short BottomOut + existing Subdialogue
[Incor + Cert]	Existing format	Existing format	Existing format if Subdialogue else new short Subdialogue + existing BottomOut

In the **Simple-adaptive experimental condition** (*Simple*), student used *Simple-adaptive* ITSPOKE-WOZ, which was discussed in Section 4.1.

In the **Complex-adaptive experimental condition** (*Complex*), student used *Complex-adaptive* ITSPOKE-WOZ, which was discussed in Section 4.2.

In the **Random-adaptive control condition** (*Random*), student used *Random-adaptive* ITSPOKE-WOZ, which provided the *Simple* adaptation to a percentage of random correct answers. This condition was included to control for the additional tutoring dialogue given to students in the two experimental conditions. The percentage was toggled during the experiment to be similar to the percentage of uncertain answers across the experimental conditions.

Fig. 10 illustrates how the system responses differ across conditions. The figure begins with an initial tutor question (**TUTOR1**) and a [Cor + Uncert] student answer (**STUDENT1**), and is then followed by four corpus excerpts.

In the *NonAdapt* excerpt, **TUTOR2** provides correctness feedback and then moves on to the next top-level question. In the *Simple* excerpt, **TUTOR2** provides correctness feedback and then initiates a Subdialogue to remediate the student's uncertainty. After this subdialogue completes, ITSPOKE move on to the next top-level question. In the *Complex* excerpt, **TUTOR2** provides feedback acknowledging both the correctness and uncertainty of the answer, then provides a BottomOut to remediate the student's uncertainty, and then moves on to the next top-level question. In the *Random* excerpt, **TUTOR2** is identical to *NonAdapt* because **STUDENT1** was not among the correct answers randomly selected to receive the *Simple* adaptation. That is, "Randomly Selected = NO". If "Randomly Selected = YES", then **TUTOR2** in *Random* and *Simple* would be identical.

Subjects were randomly assigned to the 4 conditions, except that conditions were gender-balanced and pretest-balanced. To achieve these balances, we kept a running average for pretest score and total males in each condition; we then assigned subjects to conditions based on which assignment would keep the averages similar across all conditions. Subjects were native speakers of English who had never taken college-level physics.

The *NonAdapt* condition contains 21 subjects; the other three conditions contain 20 subjects each, yielding a total of 81 subjects. Of these 81 subjects, 49 are female and 32 are male, with 7–9 males per condition.

The experimental procedure was as follows. Each subject: (i) read a physics text introducing the concepts to be tutored *(20–40 min)*; (ii) completed a pretest of 26 multiple choice questions *(20–30 min)*; (iii) used a web/voice interface to work through five physics problems with a version of the WOZ system (depending on condition) *(30–75 min)*; (iv) completed a survey questionnaire (shown in Fig. 11 and discussed in Section 7) *(10–20 min)*; and (v) completed a posttest isomorphic to the pretest *(20–30 min)*. After the experiment, the total time length of the experiment was found to vary from 1.75 h to 3.0 h across all subjects.

The corpus resulting from this experiment is described in Section 7.

6. Evaluating the adaptations: student learning results

In this section we evaluate the impact of *Simple-adaptive* and *Complex-adaptive* ITSPOKE-WOZ on student learning. Student learning is a primary performance metric in tutoring systems.

Our analysis involves statistical comparisons of learning across conditions. Our results suggest that dynamically adapting to uncertainty with substantive content can significantly improve student learning. In particular, we ran a two-way ANOVA with condition as the between-subjects factor and repeated test measures as the within-subjects factor. The ANOVA showed a significant main effect for repeated test measure ($F(1,77) = 271.214$, $p < 0.000$), indicating that students in all conditions learned a significant amount during tutoring. There was also a significant interaction effect between condition and repeated test measure ($F(3,77) = 3.275$, $p = 0.025$), indicating that how much students learned was dependent on condition.

To determine which conditions learned more, we compared two measures of learning gain: raw (posttest − pretest) and normalized ((posttest − pretest)/(1 − pretest)). For completeness, we also compared posttest score; however posttest score in isolation is less useful in our data because pretest and posttest score are highly correlated in this data ($R = 0.528$, $p < 0.000$).

For each metric, we ran a one-way ANOVA with condition as the between-subjects factor and used Tukey tests for post-hoc pairwise comparison of conditions. The ANOVAs revealed significant differences between

S1. It was easy to learn from the tutor.
S2. The tutor didn't interfere with my understanding of the content.
S3. The tutor believed I was knowledgeable.
S4. The tutor was useful.
S5. The tutor was effective on conveying ideas.
S6. The tutor was precise in providing advice.
S7. The tutor helped me to concentrate.
S8. The tutor responded effectively after I was incorrect about the answer to a question.
S9. The tutor responded effectively after I was correct about the answer to a question.
S10. The tutor responded effectively after I was uncertain about the answer to a question.
S11. The tutor responded effectively after I was certain about the answer to a question.
S12. The tutor's responses decreased my uncertainty about my understanding of the content.
S13. It was easy to understand the tutor.
S14. I knew what I could say or do at each point in the conversations with the tutor.
S15. The tutor worked the way I expected it to.
S16. Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.

5: ALMOST ALWAYS, 4: OFTEN, 3: SOMETIMES, 2: RARELY, 1: ALMOST NEVER

Fig. 11. ITSPOKE-WOZ survey questionnaire.

Table 5
Differences in student learning-related metrics across condition.

Metric	Condition	Mean	SD	Diff.	p
Raw gain	*NonAdapt*	0.183	0.108	< *Simple*	0.029
	Simple	0.307	0.127	–	
	Complex	0.213	0.114	–	
	Random	0.269	0.171	–	
Normalized gain	*NonAdapt*	0.382	0.204	< *Simple*	0.011
	Simple	0.626	0.193	–	
	Complex	0.409	0.213	< *Simple*	0.034
	Random	0.548	0.296	–	
Posttest	*NonAdapt*	0.698	0.143	< *Simple*	0.035
	Simple	0.810	0.109	–	
	Complex	0.706	0.129	–	
	Random	0.777	0.134	–	
Pretest	*NonAdapt*	0.515	0.151	–	
	Random	0.508	0.155	–	
	Simple	0.510	0.145	–	
	Complex	0.492	0.145	–	

conditions in both measures of learning gain: raw ($F(3,77) = 3.275$, $p = 0.025$) and normalized ($F(3,77) = 4.658$, $p = 0.005$), as well as for posttest score ($F(3,77) = 3.599$, $p = 0.017$). Table 5 shows the significant results ($p \leqslant 0.05$) of the pairwise comparisons for these 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, the *Simple* condition had significantly higher raw gain, normalized gain, and posttest score than the *NonAdapt* condition. The *Simple* condition also had significantly higher normalized gain than the *Complex* condition.

A comparison of the means in Table 5 shows that the *Complex* condition had higher learning gains and posttest scores than the *NonAdapt* condition, but lower learning gains and posttest scores than the *Random* condition. However, these differences are not significant.

As shown last in Table 5, we compared pretest scores across condition to determine whether or not the randomization of students in the experimental procedure was successful. We found that pretest score did not differ significantly across conditions ($F(3, 77) = 0.085$, $p = 0.968$), indicating that conditions were well-balanced for pretest score.

In sum, our results show that student learning is significantly improved by using *Simple-adaptive* ITSPOKE-WOZ to adapt to student uncertainty as compared to using the non-adaptive ITSPOKE-WOZ, but student learning is not significantly improved by using *Complex-adaptive* ITSPOKE-WOZ.

We hypothesize three possible reasons for why the *Complex* condition did not outperform any other condition significantly. First, the tutor turns may have been too long in *Complex-adaptive* ITSPOKE-WOZ; they were longer in the *Complex* condition than in all other conditions. This is illustrated in Table 7 below (Section 7, see AV # Tutor Words/Turn) and in Figs. 10 and 7–9. This increased length is due both to the longer feedback and to the use of new BottomOuts after [Cor + Uncert] and [Incor + Uncert] answers. A pilot study prior to this experiment revealed that the tutor turns in *Complex-adaptive* ITSPOKE-WOZ "feel" overly long to some users, especially to those who read them in the student interface while listening to them. It may be that these users lose focus during, and thus do not learn as much from, long tutor turns.

Second, our evaluation of *Complex-adaptive* ITSPOKE-WOZ cannot tease apart the impact of the feedback variation and the dialogue act variation of the substantive response content. In other words, it may be that these two aspects of the *Complex* adaptation had differing effectiveness overall or for specific user populations. For example, the wizard observed during the experiment that while some students seemed to like the empathetic feedback used in *Complex-adaptive* ITSPOKE-WOZ, other students seemed annoyed by it, perhaps because the system did not respond to any other affective state besides uncertainty. Thus, we cannot conclude that responding differently to different learning impasse severities (*Complex*) is less effective than treating them all the same (*Simple*).

Third, although it is common for dialogue system behavior, and particularly tutoring dialogue system behavior, to be modeled on human behavior, our results show that effective computer affect adaptations need not necessarily be isomorphic to human affect responses. In fact, it may be that different behaviors are actually optimally effective in computer and human tutors. This hypothesis is supported by our prior research, which has shown that although our students learn significantly from both our human tutor and ITSPOKE, their behaviors are very different (Forbes-Riley and Litman, 2008; Litman and Forbes-Riley, 2006b). However, we do not want to conclude that human tutor-based affect adaptations are less effective in general, because although our *Complex* adaptation was derived from statistical generalizations about human tutor responses to uncertainty, the effectiveness of these responses was not empirically tested before implementation. We return to this issue in Section 8.

Finally, we hypothesize two reasons for why the *Simple* condition did not significantly outperform the *Random* condition. For one thing, some of the turns that received the uncertainty adaptation in the *Random* condition were [Cor + Uncert] turns (13.3%), thus diminishing the difference with the *Simple* condition. This is quantified in Table 7 below (Section 7, see the last row). In our next experiment (see Section 8) we will modify the *Random* condition so that it only randomly adapts to [Cor + Cert] turns, to avoid this problem. However, it may also be the case that adapting to [Cor + Cert] answers can benefit student learning, by increasing the certainty of those answers (e.g., a [Cor + Cert] answer may be actually be neutral for certainty rather than strongly certain). On the other hand, note that although the *Random* condition adapted to more correct answers overall (19.6%) than than the *Simple* condition (10.9%) (see the penultimate row of Table 7 below), it did not yield higher learning gains. If it were equally effective to adapt to [Cor + Uncert] and [Cor + Cert] answers, then we would expect to see the *Random* condition achieving higher learning gains than the *Simple* condition, because it adapted to more correct answers overall. The fact that we do not see this suggests that it

would not further benefit learning to simply give the *Simple* adaptation to all correct answers (i.e., to give an identical response to all correct and incorrect answers (except for correctness feedback)).

7. Corpus description and wider use

The corpus resulting from our experiment consists of 405 digitally recorded (.ogg format) dialogues from 81 students, totaling approximately 70 hours of dialogue. The accompanying log files for each dialogue contain the tutor turn text (which was sent to the text-to-speech synthesis), the uncertainty annotations of the student turns (labeled by the wizard), and the student turn transcriptions, which were transcribed manually after the experiment and merged into the log files. The student transcriptions include the turn text and turn start and end times relative to the start of the dialogue, as well as punctuation and annotation of disfluencies (e.g., false starts) and non-syntactic questions (i.e., the "??" such as shown in Figs. 6–8).

Table 6 provides overall corpus details, while Table 7 shows various turn attributes broken down across the 4 conditions of the experiment.

In Table 7, the first two rows describe tutor turn attributes, while the remaining rows describe the student turns. The last two rows show how the four conditions of the experiment differed in terms of the number and type of student turns that received an uncertainty adaptation. In particular, the penultimate row shows the number and percentage of student turns that received an uncertainty adaptation, while the last row shows the percentage of adapted-to turns that were [Cor + Uncert]. For example, no turns received an adaptation in the *NonAdapt* condition. In the *Random* condition, 19.6% of turns received the *Simple* adaptation; 13.3% of these were [Cor + Uncert] turns. Since only correct turns were adapted to in the *Random* condition, 86.7% of the adapted-to turns were [Cor + Cert]. In the *Simple* condition, 10.9% of the student turns received the *Simple* adaptation; all of these were [Cor + Uncert] turns. The *Complex* condition shows the highest percent of turns adapted to (30.2%) because [Cor + Uncert], [Incor + Uncert], and [Incor + Cert] turns all received the *Complex* adaptation.

7.1. Evaluating dialogue performance

In addition to evaluating the utility of our uncertainty adaptations by comparing student learning across conditions, we are also conducting comparisons of dialogue performance across conditions as part of our ongoing work, using standard metrics from the dialogue evaluation community (e.g., Walker et al., 1997b; Möller, 2005; Forbes-Riley and Litman, 2006; Bonneau-Maynard et al., 2000). For although some of our conditions yielded no significant differences in learning, they may still have advantages due to other performance criteria.

For example, dialogue efficiency is an important performance metric for most dialogue systems, and can be measured in terms of how much time, or how many words or turns, a given task takes to complete. Dialogue efficiency is important in tutoring systems too, because students and teachers may not want to use a system that is inefficient in terms of time to task completion. While to date we've found no difference between conditions in terms of standard dialogue efficiency metrics, we have found differences with respect to learning efficiency, which is a related tutoring metric that refers to the amount of learning achieved in a given amount of

Table 6
Uncertainty corpus features.

	Student ($N = 81$)	Tutor
Total words	27,457	322,092
Total turns	6561	6561
Total uncertain turns	1491	–
Total correct turns	5147	–
Total [Cor + Cert] turns	4420	–
Total [Cor + Uncert] turns	727	–
Total [Incor + Uncert] turns	764	–
Total [Incor + Cert] turns	650	–

Table 7
Turn attributes across conditions.

Attribute	*NonAdapt*($N = 21$)	*Random* ($N = 20$)	*Simple*($N = 20$)	*Complex*($N = 20$)
AV # Tutor turns	78.0	84.5	82.2	79.5
AV # Tutor words/turn	46.6	49.5	47.6	52.2
AV # Student turns	78.0	84.5	82.2	79.5
AV # Student words/turn	5.9	4.0	3.6	3.1
AV #/% Uncertain turns	20.4/26.3	18.1/21.22	19.0/22.5	16.1/20.1
AV #/% Correct turns	58.5/75.6	66.6/79.3	65.6/80.2	63.8/80.8
AV #/% [Cor + Cert] turns	48.8/62.8	58.3/69.6	56.5/69.4	55.0/69.8
AV #/% [Cor + Uncert] turns	9.7/12.8	8.3/9.7	9.2/10.9	8.8/11.0
AV #/% [Incor + Uncert] turns	10.7/13.5	9.9/11.5	9.8/11.6	7.4/9.1
AV #/% [Incor + Cert] turns	8.8/10.9	8.1/9.2	6.8/8.2	8.4/10.1
AV #/% Turns given Uncertainty adaptation	0/0	16.5/19.6	9.2/10.9	24.5/30.2
Of Adapted-to turns, AV% that are [Cor + Uncert]	0	13.3	100.0	36.0

tutoring time (Ringenberg and VanLehn, 2006). In Forbes-Riley and Litman (2009a) we measure learning efficiency in two ways: as normalized learning gain divided by total dialogue time and by total student turns. We show that *Simple-adaptive* ITSPOKE-WOZ outperforms non-adaptive ITSPOKE-WOZ and *Complex-adaptive* ITSPOKE-WOZ on both learning efficiency metrics. These results suggest that given the same amount of tutoring time, students will learn the most from *Simple-adaptive* ITSPOKE-WOZ.

In addition, many spoken dialogue systems are evaluated in terms of a subjective user satisfaction metric, which is usually measured via a survey questionnaire and encompasses subjective perceptions of likability, ease of use, text-to-speech quality, etc. This metric is important in tutoring systems too, as students would not want to use the system if they do not like it or feel it is sufficiently usable. For this study, we constructed a survey with the statements in Fig. 11. Students rated their degree of agreement with each statement on a scale of 1–5, as shown at the bottom of the figure. Statements 1–7, taken from Baylor et al. (2003), were tailored to the tutoring domain. Statements 8–12 were tailored specifically to the uncertainty adaptations investigated in this experiment. Statements 13–16, taken from Walker et al. (2001), were more generally applicable to spoken dialogue systems. We have used statements 1–7 and 13–16 for our prior ITSPOKE corpora (Forbes-Riley et al., 2006).

In Forbes-Riley and Litman (2009a) we show that *Complex-adaptive* ITSPOKE-WOZ outperforms *Simple-adaptive* ITSPOKE-WOZ in terms of user satisfaction, in particular, with respect to student perception of tutor response quality as represented by Statement **S13**. We hypothesize that this result may indicate a student preference for the more empathetic feedback in *Complex-adaptive* ITSPOKE-WOZ.

In Forbes-Riley and Litman (2009b) we show that our learning and user satisfaction results differ for different user populations. In particular, females both learn best from and prefer *Simple-adaptive* ITSPOKE-WOZ, while males prefer *Complex-adaptive* ITSPOKE-WOZ but don't learn more from either adaptive system. Lower domain expertise users prefer *Complex-adaptive* ITSPOKE-WOZ but learn more from *Simple-adaptive* ITSPOKE-WOZ, while higher domain expertise users learn more from *Simple-adaptive* ITSPOKE-WOZ but do not prefer either adaptive system. Based on these results, we hypothesize that our uncertainty-adaptive system can be further improved by adapting to user uncertainty differently based on user classes such as gender and domain expertise.

Dialogue quality is also an important metric that can be measured in a variety of ways, including speech recognition quality (e.g., word error rate) for fully automated systems, or in terms of system-user interactivity (e.g., ratio of user to system words or turns, or average words per turn) for both fully automated and WOZ systems. Dialogue quality is important in tutoring systems too, because students and teachers may not want to use a system that doesn't permit a sufficient degree of interactivity. As shown above in Table 7, we have already computed some measures of dialogue quality for our corpus and plan to compare these and other metrics across condition in future work.

Finally, in recent work we have also investigated metrics related to metacognitive performance. In particular, we have used our wizard uncertainty annotations to compute measures of "knowledge monitoring" accuracy and related metrics (Nietfeld et al., 2006; Saadawi et al., 2009), which quantify the proportion of student answers that are actually correct when the student is judged to be certain of his/her correctness. We show that higher knowledge monitoring accuracy is predictive of higher learning in our data and that our uncertainty-adaptive systems yielded higher knowledge monitoring accuracy than our non-adaptive systems (Litman and Forbes-Riley, 2009a,b). Based on these results, we hypothesize that responding to student uncertainty can improve both cognitive and metacognitive performance, and that finding ways to improve knowledge monitoring accuracy can improve student learning. We believe that knowledge monitoring accuracy can also be a relevant construct for other dialogue applications involving knowledge asymmetry, such as problem-solving, instruction giving, and trouble shooting (e.g., Janarthanam and Lemon, 2008).

7.2. Linguistic resource for studying affect

The collected corpus itself is another important result of this study, in that it provides a novel resource for analyzing prosody and other linguistic features of naturally occurring user affect in human–computer interaction, particularly for use in automatic affect detection in spoken dialogue systems. For although there has been significant prior research on the prosody of elicited or acted emotions (e.g., Oudeyer, 2002; Liscombe et al., 2003), these results generally transfer poorly to naturally occurring emotions (Cowie and Cornelius, 2003; Batliner et al., 2003). Thus recent research on affect-adaptive spoken dialogue systems has focused on analyzing and detecting naturally occurring user affect (e.g., Vidrascu and Devillers, 2005; Batliner et al., 2003; Shafran et al., 2003).

Although this research could be substantially aided by studying affect annotated dialogues between users and adaptive systems, to date only a few such corpora have been reported or made publicly available to the computer speech and language community. For example, the HUMAINE project[4] contains a substantial collection of publicly available speech corpora annotated for speaker affective states, but very few of these corpora contain naturally occurring human–computer dialogues (e.g., Batliner et al., 2004; Walker et al., 2001; Ang et al., 2002). Moreover, of these, only the DARPA Communicator corpus uses English; it contains dialogues in the travel-planning (i.e., form-filling) domain, and user turns are annotated for the affective states of frustration and annoyance.

This corpus provides an additional resource for this active research area, because it reflects a new and complex human–computer interaction domain, it provides manual annotation of a new affective state, and it makes available a large number of features derived from the speech files and log files. We have already shown that useful predictive models of student affect in general, and student uncertainty specifically, can be built using similar features available in our prior ITSPOKE corpora (Litman and Forbes-Riley, 2006a; Ai et al., 2006). Moreover, we have observed informally that other student affective states also occur with some regularity in our corpus, namely disengagement as indicated by expressions of irritation, anger, annoyance, boredom, and humor. Although our corpus will eventually be linked to a website for access by the wider computer speech and language community, in the meantime interested users should contact the authors. In addition, we have already made publicly available a small corpus from a related pilot study (Forbes-Riley et al., 2008a,b).[5]

8. Conclusions and current directions

This article discussed the design of two versions of a wizarded spoken dialogue tutoring system that dynamically adapts to student uncertainty over and above correctness, and described a controlled experiment evaluating these two different uncertainty-adaptive systems. Both uncertainty adaptations were derived from prior tutoring research and were based on the view that uncertainty and incorrectness signal learning impasses. Both adaptive systems responded to uncertain student turns with the same substantive content that was already given to

[4] http://emotion-research.net.

[5] This pilot corpus contains approximately 20 hours of recorded dialogue and can be obtained from the Pittsburgh Science of Learning Center's Datashop at https://learnlab.web.cmu.edu/datashop/index.jsp.

incorrect turns in the original non-adaptive system, but the two adaptive systems differed in complexity. *Simple-adaptive* ITSPOKE-WOZ gave the substantive response along with correctness feedback to all learning impasses ([Cor + Uncert], [Incor + Uncert] or [Incor + Cert] answers). *Complex-adaptive* ITSPOKE-WOZ varied the dialogue act presentation of the substantive response depending on the learning impasse type, where the variations were based on statistical analysis of human tutor dialogue act variations. *Complex-adaptive* ITSPOKE-WOZ also varied its empathetic feedback phrases depending on the learning impasse type.

Our results showed that the *Simple-adaptive* ITSPOKE-WOZ significantly improved student learning as compared to the non-adaptive ITSPOKE-WOZ and *Complex-adaptive* ITSPOKE-WOZ. To our knowledge we are the first study to show that dynamically responding to student uncertainty can significantly improve learning during computer tutoring.

There were three main limitations to our study. Two of these limitations involve the design of the dialogue strategies that we developed to adapt dynamically to student uncertainty. First, although our human-tutor based *Complex* adaptation was derived from statistical generalizations about a successful human tutor's responses to uncertainty, the effectiveness of these responses was not empirically tested before implementation. In other words, although we know that students learned from our human tutor, we do not know whether or not the specific human tutor responses we found to be associated with uncertain answers are also associated with increased learning. In future work we will investigate approaches for isolating human tutor responses to uncertainty that do optimize learning. Such approaches include investigating correlations between human tutor responses to uncertainty and learning, as well as using reinforcement learning. In the larger field of spoken dialogue systems, reinforcement learning is often used to automatically extract effective dialogue system responses that depend on the current user state (Singh et al., 2002; Walker et al., 1998). In prior work with the ITSPOKE system, Tetreault and Litman (2008) used reinforcement learning to extract system responses that depended on user uncertainty as well as other user turn features. However, these responses were not implemented or evaluated in a controlled experiment. Moreover, since they were extracted from existing ITSPOKE corpora, only existing ITSPOKE responses were considered as possible adaptive strategies. In future work we will investigate using reinforcement learning in our human-tutoring corpus to determine effective human tutor responses to uncertainty. Investigating the responses of multiple human tutors may help to yield a wider variation in human tutor responses to uncertainty to consider for implementation (Porayska-Pomsta et al., 2008; Lehman et al., 2008). However, studying multiple human tutors does not necessarily yield consistent generalizations about the "best" human-based adaptive strategies to implement, because human tutors have different teaching styles and skill levels (Porayska-Pomsta et al., 2008).

Another limitation of our study is that our dialogue strategies for responding to student uncertainty were hand-crafted, rather than being generated automatically using natural language generation; thus neither the content nor the form of our uncertainty adaptations can be easily modified using parameters as would be provided by a natural language generation engine. Although currently outside the scope of our tutoring system research, the incorporation of affect into natural language generation is a very active research area in the larger field of dialogue systems, which aims to provide the technical underpinnings of fully generative affect-adaptive dialogue systems (Mairesse and Walker, 2008; Oberlander and Nowson, 2006; Walker et al., 1997a).

Finally, in this study a human wizard performed speech recognition and understanding, and uncertainty detection, thereby eliminating the potential for system errors on these tasks to impact the effectiveness of the uncertainty adaptations. Our next step will be to run a similar experiment with fully automated ITSPOKE versions replacing the WOZ system versions. In this next experiment, student uncertainty will be automatically detected, and speech recognition and natural language understanding will be fully automatic. Just as in this paper we showed that detecting and dynamically responding to student uncertainty can significantly improve student learning when these tasks are wizarded, we hope that the next experiment will show that similar results hold even when these tasks are fully automated.

Acknowledgments

This work is supported by the National Science Foundation (award number 0631930). We thank the ITSPOKE Group for help with the design and implementation of the affect-adaptive systems, and the collection of this corpus.

References

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

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 Conference Workshop on Empirical Methods for Tutorial Dialogue Systems.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Hansen, J.H.L., Pellom, B. (Eds.), Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, USA, pp. 2037–2039.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Communication 40 (1–2), 117–143.

Batliner, A., Hacker, C., Steidl, S., Noth, E., Haas, J., 2004. From emotion to interaction: lessons from real human–machine dialogues. In: Andre, E., Dybkjær, L., Minker, W., Heisterkamp, P. (Eds.), Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop, Lecture Notes in Artificial Intelligence, vol. 3068. Springer-Verlag, Berlin, pp. 1–12.

Batliner, A., Steidl, S., Hacker, C., Noth, E., 2008. Private emotions vs. social interaction-a data-driven approach towards analysing emotion in speech. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 18, 175–206.

Baylor, A.L., Ryu, J., Shen, E., 2003. The effect of pedagogical agent voice and animation on learning, motivation, and perceived persona. In: Proceedings of ED-MEDIA, Honolulu, Hawaii (June).

Bhatt, K., Evens, M., Argamon, S., 2004. Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In: Proceedings of Cognitive Science (CogSci), Chicago, USA, pp. 114–119.

Bonneau-Maynard, H. Devillers, L., Rosset, S., 2000. Predictive performance of dialog systems. In: Proceedings of Language Resources and Evaluation Conference (LREC), Athens, Greece.

Chu-Carroll, J., Nickerson, J.S., 2000. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In: Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 202–209.

Conati, C., Maclaren, H., 2004. Evaluating a probabilistic model of student affect. In: Proceedings of Intelligent Tutoring Systems Conference(ITS), Maceio, Brazil, pp. 55–66.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40 (1–2), 5–32.

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.

de Vicente, A., Pain, H., 2002. Informing the detection of the students' motivational state: an empirical study. In: Proceedings of the Intelligent Tutoring Systems Conference (ITS), Biarritz, France, pp. 933–943.

D'Mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B., Graesser, A., 2008. Automatic detection of learner's affect from conversational cues. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 18, 45–80.

Ekman, P., Friesen, W.V., 1978. The Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto.

Forbes-Riley, K., Litman, D., 2006. Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In: Proceedings Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL).

Forbes-Riley, K., Litman, D., 2007. Investigating human tutor responses to student uncertainty for adaptive system development. In: Ana Paiva, R.P., Picard, R. (Eds.), Proceedings of Affective Computing and Intelligent Interaction (ACII). Springer Press, pp. 678–689.

Forbes-Riley, K., Litman, D., 2008. Analyzing dependencies between student certainness states and tutor responses in a spoken dialogue corpus. In: Dybkjaer, L., Minker, W. (Eds.), Recent Trends in Discourse and Dialogue. Springer, pp. 275–304.

Forbes-Riley, K., Litman, D., 2009a. Adapting to student uncertainty improves tutoring dialogues. In: Proceedings 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK (July).

Forbes-Riley, K., Litman, D., 2009b. A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In: Proceedings Interspeech, Brighton, UK (September).

Forbes-Riley, K., Litman, D., Silliman, S., Tetreault, J., 2006. Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. In: Proceedings of FLAIRS.

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., Litman, D., Silliman, S., Purandare, A., 2008b. Uncertainty corpus: resource to study user affect in complex spoken dialogue systems. In: Proceedings 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco (May).

Forbes-Riley, K., Rotaru, M., Litman, D., 2008c. 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.

Graesser, A.C., Olde, B., 2003. How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. Journal of Educational Psychology 95, 524–536.

Graesser, A., Person, N., Magliano, J., 1995. Collaborative dialog patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology 9, 495–522.

Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A., 2005. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education 48 (4), 612–618.

Gratch, J., Marsella, S., 2003. Fight the way you train: the role and limits of emotions in training for combat. The Brown Journal of World Affairs 10 (1), 63–76.

Hall, L. Woods, S., Sobral, D., Paiva, A., Dautenhahn, K., Wolke, D., Newall, L., 2004. Designing empathic agents: adults vs. kids. In: Proceedings of the Intelligent Tutoring Systems Conference (ITS), Maceio, Brazil, pp. 604–613.

Janarthanam, S., Lemon, O., 2008. User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems. In: Proceedings of SEMdial.

Klein, J., Moon, Y., Picard, R., 2002. This computer responds to user frustration: theory, design, and results. Interacting with Computers 14, 119–140.

Kort, B. Reilly, R., Picard, R., 2001. An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: Okamoto, T., Hartley, R., Kinshuk, J., Klus, P. (Eds), Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges, Madison, WI, pp. 43–48.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lee, C.M., Narayanan, S., 2005. Towards detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13 (2).

Lehman, B., Matthews, M. D'Mello, S., Person, N., 2008. What are you feeling? Investigating student affective states during expert human tutoring sessions. In: Proceedings of Intelligent Tutoring Systems Conference (ITS), Montreal, Canada (June).

Liscombe, J. Venditti, J., Hirschberg, J., 2003. Classifying subject ratings of emotional speech using acoustic features. In: Proceedings of Interspeech/EuroSpeech, Geneva, Switzerland, pp. 725–728.

Litman, D., Forbes-Riley, K., 2004. Annotating student emotional states in spoken tutoring dialogues. In: Proceedings of the SIGdial Workshop on Discourse and Dialogue, Boston, USA, pp. 144–153.

Litman, D., Forbes-Riley, K., 2006a. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication 48 (5), 559–590.

Litman, D.J., Forbes-Riley, K., 2006b. Correlations between dialogue acts and learning in spoken tutoring dialogues. Journal of Natural Language Engineering: Special Issue on Educational Applications 12 (2), 161–176.

Litman, D., Forbes-Riley, K., 2009a. Improving (meta)cognitive tutoring by detecting and responding to uncertainty. In: Working Notes of the Cognitive and Metacognitive Educational Systems AAAI Symposium, Arlington, VA (November).

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

Litman, D.J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction 12 (2/3), 111–137.

Liu, K., Picard, R.W., 2005. Embedded empathy in continuous, interactive health assessment. In: CHI Workshop on HCI Challenges in Health Assessment.

Mairesse, F., Walker, M., 2008. Trainable generation of big-five personality styles through data-driven parameter estimation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL), Columbus, Ohio (June).

McQuiggan, S., Mott, B., Lester, J., 2008. Modeling self-efficacy in intelligent tutoring systems: an inductive approach. User Modeling and User-Adapted Interaction (UMUAI) 18 (1–2), 81–123.

Möller, S., 2005. Parameters for quantifying the interaction with spoken dialogue telephone services. In: Proceedings of the SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal, pp. 166–177.

Narayanan, S., 2002. Towards modeling user behavior in human–machine interaction: effect of errors and emotions. In: Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-modal Human Computer Interaction, Edinburgh, Scotland.

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.

Oberlander, J., Nowson, S., 2006. Whose thumb is it anyway? classifying author personality from weblog text. In: Proceedings of the 46 = 4th Annual Meeting of the Association for Computational Linguistics (ACL), Sydney, Australia (July).

Oudeyer, P.-Y., 2002. The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Studies 59 (1–2), 157–183.

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.

Porayska-Pomsta, K., Mavrikis, M., Pain, H., 2008. Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 18, 125–173.

Prendinger, H., Ishizuka, M., 2001. Let's talk! socially intelligent agents for language conversation training. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans (Special Issue on Socially Intelligent Agents-The Human in the Loop) 31 (5), 465–471.

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

Rozin, P., Cohen, A., 2003. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of maturally occurring facial expressions of americans. Emotion 3, 68–75.

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.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), St. Thomas, US Virgin Islands, pp. 31–36.

Singh, S., Litman, D., Kearns, M., Walker, M., 2002. Optimizing dialogue management with reinforcement learning: experiments with the NJFun system. Journal of Artificial Intelligence Research 16, 105–133.

Tetreault, J.R., Litman, D.J., 2008. A reinforcement learning approach to evaluating state representations in spoken dialogue systems. Speech Communication (Special Issue on Evaluating new methods and models for advanced speech-based interactive systems) 50 (8–9), 683–696.

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

VanLehn, K., Jordan, P.W., Rosé, C.P., 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 Intelligent Tutoring Systems.

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

Vidrascu, L., Devillers, L., 2005. Detection of real-life emotions in dialogs recorded in a call center. In: Proceedings of INTERSPEECH, Lisbon, Portugal.

Walker, M.A., Cahn, J.E., Whittaker, S.J., 1997a. Improving linguistic style: social and affective bases for agent personality. In: Mller, J. (Ed.), Proceedings of Autonomous Agents.

Walker,M.A., Litman, D., Kamm, C., Abella, A., 1997b. PARADISE: a general framework for evaluating spoken dialogue agents. In: Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL), Madrid, Spain, pp. 271–280.

Walker, M., Fromer, J., Narayanan, S., 1998. Learning optimal dialogue strategies: a case study of a spoken dialogue agent for email. In: Proceedings of ACL/COLING.

Walker, M., Passonneau, R., Boland, J., 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Toulouse, France, pp. 515–522.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002. DARPA Communicator: cross-system results for the 2001 evaluation. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, USA, pp. 269–272.

Wang, N., Johnson, W., Rizzo, P., Shaw, E., Mayer, R., 2005. Experimental evaluation of polite interaction tactics for pedagogical agents. In: Proceedings of Intelligent User Interface Conference (IUI), pp. 12–19.

Wang, N., Johnson, W., Mayer, R.E., Rizzo, P., Shaw, E., Collins, H., 2008. The politeness effect: pedagogical agents and learning outcomes. International Journal of Human–Computer Studies 66 (2), 98–112.

Yannakakis, G.N., Hallam, J., Lund, H.H., 2008. Entertainment capture through heart rate activity in physical interactive playgrounds. User Modeling and User-Adapted Interaction: The Journal of Personalization Research 18, 207–243.