Human-Computer Interaction
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653648

Toward Spoken Human-Computer Tutorial Dialogues

Sidney K. D'Mello[a]; Art Graesser[a]; Brandon King[b]
[a] University of Memphis, [b] University of California, Davis

Online publication date: 15 December 2010

To cite this Article D'Mello, Sidney K. , Graesser, Art and King, Brandon(2010) 'Toward Spoken Human-Computer Tutorial Dialogues', Human-Computer Interaction, 25: 4, 289 — 323
To link to this Article: DOI: 10.1080/07370024.2010.499850
URL: http://dx.doi.org/10.1080/07370024.2010.499850

HUMAN–COMPUTER INTERACTION, 2010, Volume 25, pp. 289–323

ISSN: 0737-0024 print / 1532-7051 online
DOI: 10.1080/07370024.2010.499850

Taylor & Francis
Taylor & Francis Group

Toward Spoken Human–Computer Tutorial Dialogues

Sidney K. D'Mello,[1] Art Graesser,[1] and Brandon King[2]
[1]University of Memphis
[2]University of California, Davis

Oral discourse is the primary form of human–human communication, hence, computer interfaces that communicate via unstructured spoken dialogues will presumably provide a more efficient, meaningful, and naturalistic interaction experience. Within the context of learning environments, there are theoretical positions supporting a speech facilitation hypothesis that predicts that spoken tutorial dialogues will increase learning more than typed dialogues. We evaluated this hypothesis in an experiment where 24 participants learned computer literacy via a spoken and a typed conversation with AutoTutor, an intelligent tutoring system with conversational dialogues. The results indicated that (a) enhanced content coverage was achieved in the spoken condition; (b) learning gains for both modalities were on par and greater than a no-instruction control; (c) although speech recognition errors were unrelated to learning gains, they were linked to participants' evaluations of the tutor; (d) participants adjusted their conversational styles when speaking compared to typing; (e) semantic and statistical natural language understanding approaches to comprehending learners' responses were more resilient to speech recognition errors than syntactic and symbolic-based approaches; and (f) simulated speech recognition errors had differential impacts on the fidelity of different semantic algorithms. We discuss the impact of our findings on the speech facilitation hypothesis and on human–computer interfaces that support spoken dialogues.

Sidney D'Mello is a computer scientist with an interest in affective computing, human–computer interaction, intelligent tutoring systems, and computational modeling of human cognition; he is a Postdoctoral Researcher in the Institute for Intelligent Systems at the University of Memphis. **Art Graesser** is a psychologist with an interest in knowledge representation, question asking and answering, tutoring, text comprehension, inference generation, conversation, and reading; he is a Full Professor in the Department of Psychology at the University of Memphis. **Brandon King** is a psychology researcher with an interest in emotion regulation and cognitive control; he is a Junior Specialist at the Center for Mind and Brain at the University of California, Davis.

CONTENTS

1. INTRODUCTION
    1.1. Advantages of Spoken Tutorial Dialogues
    1.2. Research Goals
2. A BRIEF OVERVIEW OF AUTOTUTOR
    2.1. The Structure of AutoTutor Dialogue
    2.2. Interpreting Learners' Responses
3. METHODS
    3.1. Participants
    3.2. Materials
        AutoTutor
        Knowledge Tests and Evaluation Questionnaires
    3.3. Procedure
    3.4. Speech Recognition Accuracy
4. RESULTS: IMPACT OF SPOKEN INPUT ON THE STUDENT (GOALS 1–4)
    4.1. Problem Completion Rates for Spoken and Typed Conditions
    4.2. Learning Gains for Spoken and Typed Conditions
    4.3. User Satisfaction for Spoken and Typed Conditions
    4.4. Conversational Styles for Spoken and Typed Conditions
        Response Verbosity
        Answer Attempts
        Speech Acts
        Word Usage
        Relationship Between Conversational Styles and ASR Errors
5. RESULTS: EFFECT OF ASR ERRORS ON THE TUTOR (GOALS 5–6)
    5.1. Speech Act Classification
    5.2. Meaning Assessment Module
    5.3. Word Error Rate Simulations to Generalize Findings
6. GENERAL DISCUSSION
7. CONCLUSIONS

1. INTRODUCTION

The ability to have a natural language spoken conversation with a computer is a technological feat that has been long desired, but the achievements have slowly accumulated and have arguably been modest. Developing artificial spoken dialogues in natural language involves an integration of systems that combine automatic speech recognition (ASR) for speech-to-text translation with natural language processing to link the recognized text to specific computer actions. Such systems mark a significant departure from the currently dominant window, icon, menu, and pointing device (WIMP) style of interaction.

The major motivating factor behind redefining the interaction paradigm from WIMP to spoken dialogues is to narrow the communicative barrier between computers and humans. Humans primarily communicate through speech and a host of non-

verbal cues, such as facial expressions, posture, and gesture, rather than through typing and clicking. Computer systems that are able to recognize and respond to these communication channels will presumably provide a more efficient, meaningful, and naturalistic interaction experience. The ability to communicate with a computer though natural speech would represent a significant advancement toward cracking the barrier between the highly expressive human and the socially challenged computer.

One class of technologies that might greatly benefit from spoken input is intelligent tutoring systems (ITSs). These systems mimic one-on-one human tutoring, which is a powerful method of promoting active construction of knowledge beyond textbooks and traditional classroom environments (Bloom, 1984; P. Cohen, Kulik, & Kulik, 1982; Corbett, 2002; VanLehn, 2006). ITSs have implemented several systematic strategies for increasing learning gains, such as error identification and correction, building on prerequisites, frontier learning (expanding on what the learner already knows), student modeling (inferring what the student knows and having that information guide tutoring), and building coherent explanations (Aleven & Koedinger, 2002; Anderson, Corbett, Koedinger, & Pelletier, 1995; Gertner & VanLehn, 2000; Graesser, Person, Lu, Jeon, & McDaniel, 2005; Koedinger, Anderson, Hadley, & Mark, 1997; Lesgold, Lajoie, Bunzo, & Eggan, 1992; Sleeman & Brown, 1982; VanLehn, 2006). The ITSs that have been successfully implemented and tested (such as the Andes physics tutor, Cognitive Tutor, and AutoTutor) have produced learning gains of approximately 1.0 standard deviation units (sigma), or approximately one letter grade (Corbett, Anderson, Graesser, Koedinger, & VanLehn, 1999; Graesser et al., 2004; VanLehn et al., 2007). This is an impressive feat because the 1.0 sigma effect size produced by ITSs is superior to the 0.4 sigma effect size obtained from novice human tutors (P. Cohen et al., 1982), although it is lower than the 2.0 sigma effect obtained by some expert human tutors in mathematics (Bloom, 1984).

It appears that there are two different mechanisms that potentially explain the effectiveness of one-on-one tutoring (Corbett et al., 1999; Graesser, Person, & Magliano, 1995). The first is the sophisticated tutoring strategies that have been identified in the ITS literature (Psotka, Massey, & Mutter, 1988; Sleeman & Brown, 1982; Woolf, 2009). The second is the dialogue patterns and natural language that help human tutors scaffold the learner to new levels of mastery (Chi, Roy, & Hausmann, 2008; Graesser et al., 1995). According to Graesser et al. (1995), there is something about discourse and natural language (as opposed to sophisticated pedagogical strategies) that to some extent explains the effectiveness of novice human tutors. They arrive at this conclusion because most novice human tutors are effective, but they use very few if any sophisticated pedagogical strategies. Perhaps a combination of sophisticated tutoring strategies and conversational patterns will produce the ideal tutoring system.

1.1. Advantages of Spoken Tutorial Dialogues

It could be argued that spoken interaction is important in tutorial dialogue, compared to communication through typing and print (Litman et al., 2006; Pon-Barry, Clark, Schultz, Bratt, & Peters, 2004; Whittaker, 2003). One reason is that the expres-

sion gap between thought and speech is much less than the gap between thought and writing, because oral discourse is the language of the "mother tongue" (Chafe, 1982; Tannen, 1982). Hence, spoken responses are relatively effortless and more natural, at least when compared to written responses (De La Paz & Graham, 1997). Because of the ease of spoken responses, the volume of content is typically longer when spoken than typed. Given that learning is correlated with the volume of responses by the students (Chi, Siler, & Jeong, 2004), following a constructivist framework (Dalgarno, 2001; Moshman, 1982), it would be predicted that spoken responses would yield higher learning gains.

A modality hypothesis would also be compatible with the predictions of the speech-facilitation hypothesis (Mayer, 2005; Mayer, Sobko, & Mautone, 2003). The modality hypothesis states that a particular modality might get overloaded when there are multiple channels of information in the same modality. The students in the ITS we tested (AutoTutor, which is described later) has a talking head with facial expressions, diagrams on the subject matter, and three windows with text information. The visual channel is quite busy, so adding another text window for typing runs the risk of cognitive overload. For example, when answering questions involving images, students will have to constantly shift their attention from the image to the text-box when constructing a typed response. This problem can be mitigated to some extent by students speaking their responses instead of typing them. The students can focus on aspects of the image while constructing their spoken responses, thereby reducing attentional demands.

There is also a stylistic advantage of spoken communication channels in instructional design. According to social agency theory (Mayer et al., 2003), when social cues are used in computerized learning environments, learners engage in deeper cognitive processing than when these social cues are absent. Students prefer human voices over machine-synthesized voices (Mayer et al., 2003) and conversational over formal speech (Mayer, Fennell, Farmer, & Campbell, 2004). Spoken instruction may prime the social rules of human-to-human communication, resulting in pragmatically appropriate processing of incoming material by learners. Social agency theory does not predict a difference in learning gains when tutor output modality is spoken, but student input modality varies between speech and typed (like the current experiment). However, although Mayer's claims apply to students perceiving and comprehending input, the advantages of the spoken dialogues might also apply to students' production of information, which is the emphasis of this article.

The aforementioned benefits of spoken input suggest a *speech-facilitation hypothesis,* which predicts that spoken input will *increase* learning gains compared to typed input. However, the one study that tested this hypothesis produced mixed results. We are referring to Litman and colleagues' (2006) study of a spoken dialogue-based physics ITS called ITSPOKE (Litman & Silliman, 2004). They conducted two experiments that evaluated the effectiveness of spoken student–tutor dialogues along a number of dependent measures, including learning gains and time on task. The first experiment compared typed versus spoken dialogues with a human tutor, whereas the second experiment compared typed versus spoken dialogues with a computer tutor. Their re-

sults indicated that changing the input modality from typed to spoken dialogues had a substantial positive impact on learning and reduced training time in the experiment with the human tutor but had no impact on learning with the computer tutor. Although there appears to be advantages to spoken tutorial interventions, the benefits might apply only to human–human communication and not to human–computer interactions. However, the results should be interpreted with a modicum of caution because they have not been adequately replicated. Systematic replications with different students, tutors, ITSs, and topics might support alternate conclusions.

Hence, we tested the *speech-facilitation hypothesis* with a different ITS called AutoTutor on a different domain (computer literacy). AutoTutor is a fully automated computer tutor that simulates a human tutor and holds conversations with students in natural language (Graesser, Chipman, Haynes, & Olney, 2005; Graesser et al., 2004). The study involved 24 students interacting with a new spoken-input version and the traditional typed-input version of AutoTutor. There was also a no-treatment control condition in which students did not receive any tutoring.

1.2. Research Goals

The current study had six goals. The first four goals tested some of the predictions of the aforementioned theoretical perspectives that predict that input modality will have differential effects on content coverage, learning gains, learners' subjective evaluations of the tutorial session, and learners' conversational styles.

Goal 1 (Enhanced content coverage) tested whether spoken dialogues afforded enhanced content coverage compared to typed dialogues due to the efficiency of speech production (see previously). Content coverage was measured by the number of problems completed in a fixed amount of time.

Goal 2 (Speech facilitation hypothesis) evaluated whether learning gains were higher when learners spoke their responses versus typing them in. Learning gains were measured on the basis of knowledge tests administered before and after the tutorial session.

In addition to achieving learning gains, enhancing user satisfaction is another important goal of ITSs. There are reasons to expect that user satisfaction might be lower in spoken-input tutoring systems. Perhaps people are not used to speaking aloud to computers and may be uncomfortable with this activity. This potential dislike of spoken interfaces when compounded with speech recognition errors might lead to discouragement, disappointment, or even frustration. A lack of confidence in an ITS caused by speech recognition errors is likely to have a negative impact on learning. Hence, Goal 3 (Evaluations of tutorial session) was to compare learners' subjective evaluations of the spoken versus typed versions of AutoTutor. Learners' impressions of both versions of AutoTutor were measured via questionnaires administered after each tutorial session.

Goal 4 (Conversational styles) tested whether learners adjusted their conversational styles while speaking rather than typing. As social agency theory predicts, this adjustment might occur because learners recognize that the computer is attempting to

communicate with enhanced social skills, so they respond in a more social manner. It is also possible that learners are sympathetic to the system's limitations in recognizing and understanding their speech. Hence, they might adapt the manner in which they interact with a computer that supports spoken dialogues (i.e., speaking clearly, slowly, without idiomatic expressions, etc.). This goal was addressed by investigating the extent to which learners adjusted their conversational styles (i.e., different speech acts, question asking behaviors, word usage, etc.) while speaking rather than typing.

Whereas the first four goals of this research addressed the impact of spoken input and ASR errors from the perspective of the student, Goals 5 and 6 address how these errors affect the computer tutor. This is an important concern because ASR systems are imperfect, with word error rates (WERs) for conversational speech ranging from 13.5% to 45.5%, depending on the system, domain, and testing environment (Hagen, Pellom, & Cole, 2007; Kato, Nanjo, & Kawahara, 2000; Leeuwis, Federico, & Cettolo, 2003; Litman et al., 2006; Munteanu, Penn, Baecker, & Zhang, 2006; Pellom & Hacioglu, 2003; Rogina & Schaaf, 2002; Zolnay, Kocharov, Schluter, & Ney, 2007). Furthermore, several of these ASR systems have been tested in laboratory settings with negligible ambient noise sources and with simple dictation tasks, such as speakers reading news articles. A substantial degradation in performance is expected in real-world environments with extraneous sources of noise. The situation is further complicated in learning environments where students' speech is replete with pauses, uncertainty, hesitations, mispronunciations, and imperfect phrasings.

It is conceivable that ASR errors are less problematic in spoken dialogue systems if the content of the user's utterance (response) can be linked to an appropriate tutor action. In other words ASR does not have to be perfect if the intention and core ideas of the utterance are understood by the system. Hence, aside from the speech dimensions of ASR systems, ITSs with spoken dialogue require adequate natural language understanding (NLU) capabilities for comprehending the learners' responses. The last two decades have seen major advances in NLU technologies (Allen, 1995; Jurafsky & Martin, 2008), but current NLU systems are imperfect. Clearly, NLU is a formidable challenge for both spoken and text-based dialogues. However, additional complexities arise for NLU with speech, because ASR errors are expected to degrade performance of the NLU systems.

Robust ASR and NLU capabilities are the most formidable challenges that spoken-input interfaces must overcome. However, it has been shown that perfect ASR and NLU are not a requirement for functional spoken dialogue systems in tutoring. Examples of systems that are effective despite ASR and NLU imperfections include Litman's ITSPOKE system (Litman et al., 2006; Litman & Silliman, 2004), Mostow et al.'s Reading Tutor (Mostow & Aist, 2001), the Scot system (Pon-Barry et al., 2004; Schultz et al., 2003), and the Tactical Language and Culture Training System (Johnson & Valente, 2008).

Although these systems exemplify the possibility that perfect ASR and NLU are not necessary for an effective tutorial interaction, two important questions remain. There is the question of determining the appropriate level to analyze spoken utterances when the utterances are confounded with ASR errors. An important contrast to

consider is between shallow and deep NLU, which are the two broad categories of models that interpret a user's utterance. Unlike deep NLU, shallow NLU techniques do not perform a thorough linguistic analysis of the user's utterance. Shallow NLU relies on techniques such as keyword matching, matching decomposed morphemes, regular expressions, simple rule-based grammars, ngrams, and statistical approaches to semantics, syntax, and world knowledge (Jurafsky & Martin, 2008; Landauer, McNamara, Dennis, & Kintsch, 2007).

In contrast, deep NLU approaches attempt to achieve a precise, complete, well-formed analysis of the syntax, semantics, and discourse. Deep NLU requires computational linguistic tools such as syntactic parsers, part-of-speech taggers, morphological transducers, lemmatizers, reference resolution, semantic decomposition, and sometimes discourse analyzers (Jurafsky & Martin, 2008). These algorithms work best when the utterance is syntactically intact, a feat that is exceedingly difficult to achieve in spoken conversations because of the disfluencies, hesitations, and perturbations that routinely accompany natural speech (Levinson, 1983; Lickley, 1994; Shriberg, 1994).

Imperfect ASR systems undoubtedly add to the degraded quality of learners' speech. Therefore, it is conceivable that ASR errors may severely impact the functionality of deep NLU systems. This fundamentally empirical question was investigated in Goal 5 (Comparatively deep versus shallow NLP) of this research. In particular, we assessed the impact of ASR errors on AutoTutor's Speech Act Classifier, which performs a syntactic-semantic analysis of a learner's utterance in order to classify it as domain-related response (i.e., contribution) versus particular types of frozen expression. If the response is classified as a contribution, the Meaning Assessment module (or Assessment module) performs a semantic analysis to assess the conceptual quality of the response. Because the SAC module relies on comparatively deeper NLU than the Assessment module, we expect the Assessment module to be more resilient to ASR errors than the SAC module.

Although shallow NLU modules may conceivably be resilient to ASR errors, the ASR system must have a modicum of accuracy. But the degree of errors that can be tolerated before the interaction is compromised is an important, yet open research question. This question was investigated in Goal 6 (ASR error simulation) where we simulated different degrees (0% errors to 80% errors) of ASR errors and investigated their impact on the tutor's ability to comprehend learners' responses. Specifically, we compared how different degrees of ASR errors affected the performance of two NLU algorithms used to comprehend learners' responses.

The remainder of this article begins with a description of AutoTutor from the standpoint of its natural language tutorial dialogue. We then describe the methodology of the study. The results section is divided into two parts. First we investigate the impact of spoken dialogues and ASR errors from the perspective of the human students; the dependent variables are problem completion rates, learning gains, subjective evaluations of the interaction, and a host of other dialogue interaction features (Goals 1–4). The second half of the results focuses on the impact of ASR errors on AutoTutor's NLU mechanisms (Goals 5–6). We conclude by discussing the implications of this

research and the generalizability of our major findings to other ITSs and software interfaces.

2. A BRIEF OVERVIEW OF AUTOTUTOR

AutoTutor is an intelligent tutoring system that helps students learn Newtonian physics, computer literacy, and critical thinking topics through tutorial dialogue in natural language (Graesser et al., 2004; Storey, Kopp, Wiemer, Chipman, & Graesser, in press; VanLehn et al., 2007). The impact of AutoTutor has been validated in approximately 20 experiments (Graesser et al., 2004; Storey et al., in press; VanLehn et al., 2007). Hence, from the standpoint of the present study, we take it as given that the conventional AutoTutor with typed input helps learning, whereas our present focus is on the input modality (typed or spoken) of the tutorial session.

AutoTutor's dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. For example, the following is an example of a challenging question in computer literacy: *When you turn on the computer, how is the operating system first activated and loaded into RAM?* These questions require the learner to construct approximately three to seven sentences in an ideal answer and to exhibit reasoning in natural language. However, when students are asked these challenging questions, their initial answers are typically only one or two sentences in length. The one or two sentences provide insufficient information to adequately answer the question, so a tutorial dialogue is needed to flesh out a complete answer. AutoTutor engages the student in a mixed-initiative dialogue that draws out more of what the student knows and assists the student in the construction of an improved answer (see Figure 1 for an excerpt of a conversation with AutoTutor extracted from an actual tutoring session).

AutoTutor's interface has five windows shown in Figure 2. The Question window (top of screen) is the main question that stays on the computer screen throughout the conversation for the question. The Agent window (left middle) is the animated conversational agent that speaks the content of AutoTutor's turns. The Image window (right middle) either is blank or has auxiliary diagrams. The Dialogue History window (left bottom) displays the dialogue history of the tutoring session. As discussed next, this window was disabled for the current experiment. Finally, the Student Response window (right bottom) displays students' answers as they type them in.

2.1. The Structure of AutoTutor Dialogue

AutoTutor's dialogue was designed to simulate human tutoring (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser et al., 1995; Shah, Evens, Michael, & Rovick, 2002). The nature of the dialogue has been described in detail in previous publications (Graesser, Chipman, et al., 2005; Graesser et al., 2004). Expectations and misconceptions form the underlying conceptual structure that drives AutoTutor's dialogue and

FIGURE 1. Sample conversation between AutoTutor and a student.

Speaker	Text
AutoTutor	When you turn on the computer, how is the operating system first activated and loaded into RAM? [Main question]
Student	*When the computer is turned on a file is automatically booted up.*
AutoTutor	Anything else? [Pump for information]
Student	*This file contains the operating system.*
AutoTutor	What about storage? [Hint]
Student	*The operating system is then read into RAM and activated.*
AutoTutor	Right. [Positive feedback]. The CPU executes instructions in ROM, which helps the operating system to start. [Assertion]

FIGURE 2. Screenshot of the AutoTutor interface.

are the primary pedagogical methods of scaffolding good student answers. Both AutoTutor and human tutors (Graesser et al., 1995) typically have a list of expectations (anticipated good answers) and a list of anticipated misconceptions associated with each main question. For example, the following four expectations are associated with the question "Why do computers need operating systems?": (E1) The operating system helps load application programs, (E2) The operating system coordinates communications between the software and the peripherals, (E3) The operating system allows

communication between the user and the hardware, and (E4) The operating system helps the computer hardware run efficiently.

AutoTutor guides the student in articulating each of the expectations of a problem (or main question) through a five-step dialogue frame that is prevalent in human tutoring (Graesser et al., 1995; VanLehn et al., 2007). The five steps of the dialogue frame are (a) Tutor asks main question, (b) Student gives initial answer, (c) Tutor gives short feedback on the quality of the student's answer in the second step, (d) Tutor and student collaboratively interact via expectation and misconception tailored dialogue, and (e) Tutor verifies that the student understands (e.g., "Do you understand?").

This dialogue frame is implemented over a number of conversational turns. Each turn of AutoTutor in the conversational dialogue has three information slots (i.e., constituents). The first slot of most turns is short feedback on the quality of the student's last turn. This feedback is positive (e.g., "very good," "bravo"), negative (e.g., "not quite," "no"), or neutral (e.g., "uh-huh," "okay"). AutoTutor's feedback is manifested in its verbal content, intonation, and other nonverbal conversational cues. The second slot advances the coverage of the ideal answer with prompts for specific words ("X is a type of what?"), hints ("What can you say about X?"), assertions with correct information ("X is required for …")., corrections of misconceptions, or answers to students' questions (via information retrieval from a glossary or textbook). The third slot is a cue to the student that the conversation flow is being shifted from AutoTutor to the student. For example, AutoTutor ends each turn with a question or a gesture (rendered by the animated conversational agent) to cue the learner to do the talking. Discourse markers (e.g., "and also," "okay," "well") connect the utterances of these three slots of information within a turn.

As the learner expresses information over many turns, the list of expectations is eventually covered and the main question is scored as answered. Complete coverage of the answer requires AutoTutor to have a pool of hints and prompts available to extract all of the content words, phrases, and propositions in each expectation. AutoTutor selects those hints and prompts that fill missing constituents and thereby achieves pattern completion. Empirical validation of the quality of AutoTutor's dialogue has been presented in previous publications (Graesser et al., 2003; Person & Graesser, 2002).

2.2. Interpreting Learners' Responses

The structure of AutoTutor dialogue, as previously described, goes a long way in simulating a novice human tutor. AutoTutor can keep the dialogue on track because it is always comparing what the student says to anticipated input (i.e., the expectations and misconceptions in the curriculum script). Pattern matching operations and pattern completion mechanisms drive the comparison. These matching and completion operations are based on symbolic interpretation algorithms (Rus & Graesser, 2007) and statistical semantic matching algorithms (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007). A two-tiered interpretation scheme is used to understand the learner's utterance. First, a Speech Act Classifier determines whether the incoming utterance best fits a contribution (i.e., directly related to the tutoring content, e.g., "switch to virtual

memory") or a frozen expression that signifies a particular discourse function. Frozen expressions include a short response (e.g., "yes," "no"), a metacognitive statement (e.g., "I need help," "I don't know"), a metacommunicative statement (e.g., "please repeat," "could you say that in another way"), and a question (e.g., "what is RAM?"). If the response is classified as a contribution, semantic matching algorithms attempt to compute the conceptual quality of the response. Details on the actual mechanisms that AutoTutor uses to interpret the learner's contributions are presented in the subsequent sections. Evaluations of the tutor's ability to comprehend learners' responses are discussed in previous publications (Graesser et al., 2007; Graesser et al., 2000).

3. METHODS

The present study had a repeated-measures experimental design in which each participant was assigned to all three tutor conditions: spoken input, typed input, and a no-intervention control (no tutoring). The order in which students were assigned the input methods (spoken first and typed second or typed first and spoken second) was counterbalanced across participants. For each participant, these three experimental conditions were assigned to one of three topics in computer literacy: hardware, operating systems, and Internet. Assignment of the computer literacy topics to tutor conditions was counterbalanced across all participants using a 3×3 Latin square.

3.1. Participants

Twenty-four undergraduate students (10 men, 14 women) participated in the study. Fifteen of these participants were recruited from the Psychology Subject Pool at the University of Memphis and received course credit in return for their participation. The remaining nine participants were recruited through on-campus advertisement and received $20 in monetary compensation for their participation. Data from one participant were eliminated due to experimenter error.

3.2. Materials

AutoTutor

Two versions of AutoTutor tutored students on the topic of computer literacy. The typed input version was the traditional AutoTutor, where participants typed their responses into a text area and pressed the Enter key to submit their responses to the tutor. The submission pane displayed only the dialogue move that participants were currently constructing. Once a dialogue move was submitted to the tutor, it was no longer available for participants to view on the screen (as discussed next).

In the spoken input version, students spoke their responses through a microphone and were unable to see the text of their dialogue moves. Participants' speech and the audio generated by AutoTutor's animated conversational agent were recorded

for offline analyses. The participants pressed the F1 key to initiate a spoken response. Once the participants finished their spoken response, they pressed the F2 key to submit this response to the tutor.

We used the commercially available Dragon NaturallySpeaking™ ASR system (Nuance Communications, Burlington, MA) for speech-to-text translation. Prior to interacting with AutoTutor, participants completed a 7- to 10-min training period on the speech recognition system during which they read aloud a passage of text. This enabled the speech recognition system to build a profile of the participant's speech patterns to improve recognition accuracy. The text of the training passage was unrelated to the learning content of the tutorial session. To further increase recognition accuracy, the speech recognition system was equipped with a custom dictionary of 3,050 content-specific words (e.g., *RAM*, *e-mail*, *hard drive*) in addition to the commonly used words (e.g., *and*, *because*, *are*) that are included in the speech system by default.

It is important to highlight four additional points about the two versions of AutoTutor. First, the history of the dialogue between the student and the tutor (see bottom left window in Figure 2) was not available for the participants to view in both conditions. The rationale behind this decision was to prevent participants from viewing the (error-ridden) output of the speech recognizer. Second, in both versions, an embodied pedagogical agent (see middle left of Figure 2) delivered all of the tutor's dialogue via synthesized speech. For the typed version, the tutor spoke and the learners typed their responses. For the spoken version, both the student and the tutor spoke their responses. Thus, any detected differences can be attributed to students input modality and not to the tutor's output modality. Third, due to the increased processing time required for analyzing ASR input, the spoken-input version of the tutor took longer (< 6 s) to respond. Although response times for the typed version were in the 1- to 3-s range, follow-up analyses revealed that the increased response latency of the spoken version had no impact on problem completion times. Finally, unlike the Litman et al. (2006) human tutoring experiment, where strict turn taking was enforced in the typed condition but not the spoken condition (a potential confound), strict turn taking was enforced in both conditions in the current experiment.

Knowledge Tests and Evaluation Questionnaires

Participants were tested on their knowledge of computer literacy topics both before and after the tutorial session (pretest and posttest, respectively). The testing materials were adapted from computer literacy tests used in previous experiments involving AutoTutor (Graesser et al., 2004). They comprised four-alternative multiple-choice questions that assessed students' knowledge of all three computer literacy topics. Each test contained 24 multiple-choice questions with eight questions on each of the three topics. There were two test versions that tested learners on the same subject matter and content but were composed of different questions. The assignment of test versions to pretest versus posttest was counterbalanced across participants.

The four-alternative multiple-choice format was designed to assess deep levels of knowledge. The questions required answers that involved inferences and deep reason-

ing, such as *why, how, what-if, what if not, how is X similar to Y?* These questions that assess deep levels of knowledge can be contrasted with those that assess shallow levels of knowledge by simply asking students to recall previously presented information, definitions, and facts (Graesser, Ozuru, & Sullins, 2010). As an example, consider the following hardware question: "If you install a sound card, why does your computer perform better?" Alternatives for this question were (a) Because it can bypass the operating system when sound is needed, (b) Because it does not need the CPU, (c) Because sound no longer requires RAM, and (d) Because there will be fewer bottlenecks when multitasking.

The eight-item postinteraction questionnaire asked participants to evaluate their tutorial session on measures of perceived performance, user satisfaction, and task difficulty. The first six questions required participants to rate the following statements on a 6-point scale (*strongly agree* to *strongly disagree*): (a) "I enjoyed interacting with AutoTutor," (b) "I felt that my interaction with AutoTutor was comparable to an interaction with a human tutor," (c) "I felt that AutoTutor did *not* understand what I said," (d) "I felt engaged during the tutoring session," (e) "I felt that AutoTutor was difficult to use and work with," and (f) "I felt that I learned new information from AutoTutor."

The questionnaire also contained two items designed to assess participants' perceived intensity of mental effort, which is a reliable measure of cognitive load during learning (Paas, van Merrienboer, & Adam, 1994). These items were adapted from a survey used to assess perceived cognitive load during a multimedia presentation (Mayer et al., 2003). The first item asked participants to rate the difficulty of learning computer literacy from the tutor on a 7-point scale (*very easy* to *very hard*): "How difficult was it for you to learn about computer literacy from the tutoring session you just participated in?" The second item asked participants to rate the difficulty of the session independently of the learning content on the same 7-point scale: "Apart from the content of the tutoring session, how difficult was it to learn new information from the tutor?"

3.3. Procedure

Participants were tested individually during a 2-hr session. First, participants completed an informed consent and then the pretest. Next, the general features of AutoTutor's dialogue and pedagogical strategies were described to the participants. Participants then interacted with one version of AutoTutor until four questions were successfully answered or 35 min had elapsed. After the tutorial interaction, participants completed the postinteraction questionnaire (on paper). The tutorial interaction and postinteraction questionnaire were then repeated for the second version of AutoTutor. Finally, participants in both groups completed the posttest and were debriefed.

3.4. Speech Recognition Accuracy

The tutorial session yielded 1,061 student–tutor conversational turns ($M = 46$ per student, $SD = 8$) for the spoken condition and 987 turns ($M = 43$, $SD = 6$) for the typed condition. To evaluate the accuracy of the speech recognition system, the automatically recognized speech of each student turn was compared to a manual transcrip-

tion of the speech in that turn, which was prepared by an experimenter from the audio recordings of the tutorial interactions.

The mean word recognition rate[1] (WRR) for our ASR system was .542 (SD = .270, min = .017, max = .887, median = .595), which is moderate accuracy for automatic recognition of natural conversational speech. We performed an analysis on the proportions of substitution, deletion, and insertion errors in each turn. A 3 × 3 × 2 repeated measures analysis of variance (ANOVA) was conducted with *error type* (substitutions, deletions, insertions) as a within-subjects factor and *tutoring topic* (hardware, internet, operating systems) and *interaction order*[2] (spoken-typed, typed-spoken) as between-subjects factors. The only significant[3] effect was the main effect for error type, $F(2, 34) = 37.82$, $MSe = .03$. Bonferroni post hoc tests confirmed that the proportion of substitution ($M = .523$, $SD = .141$) and insertion errors ($M = .391$, $SD = .173$) did not differ and were significantly higher than deletion errors ($M = .088$, $SD = .053$).

When word order is ignored, 75% of the words were correctly recognized (SD = 15.3%). These results suggest that our ASR system would be problematic for a system that requires a syntactically intact utterance to evaluate a learner's response. However, performance is expected to be relatively stable for ITSs that rely on shallow NLU techniques because these algorithms match key words and phrases while ignoring the syntax of the utterance.

4. RESULTS: IMPACT OF SPOKEN INPUT ON THE STUDENT (GOALS 1–4)

In accordance with the first four goals of this research, this set of analyses investigated the impact of input modality from the perspective of the human student. Specifically, we investigated four questions that contrasted spoken and typed input from the perspectives of (a) problem completion rates, (b) learning gains, (c) participants' evaluations of the tutorial session, and (d) participants' conversational style. This section addresses these questions as well as assesses the correlation of ASR errors with each of these dependent variables.

4.1. Problem Completion Rates for Spoken and Typed Conditions

We compared problem completion rates to test whether spoken dialogues afforded enhanced content coverage compared to typed dialogues due to the efficiency of speech production (Goal 1). We coded problem completion rates on a

[1]WER and WRR are standard metrics for assessing the reliability of automatic speech recognition systems. WER = $[S + D + I]/N$, where S, D, and I are the number of substitutions, deletions, and insertions in the automatically recognized text (with errors) when compared to the ideal text (no errors) of N words. $WRR = 1 - WER$.

[2]In this and subsequent analyses, the interaction order (spoken then typed vs. typed then spoken) was included as a between-subjects factor. However, the main effect for interaction order was never significant, nor were the two-way interactions between interaction order and other variables. Therefore, speaking first and then typing, versus typing first and then speaking, had no impact on the dependent variables.

[3]$p < .05$ in this and subsequent analyses unless explicitly noted.

5-point scale—0.0, 0.25, 0.5, 0.75, and 1.0 for 0, 1, 2, 3, and 4 problems completed in the 35-min time limit. A repeated measures ANOVA revealed that there was a significant difference in problems completed across conditions, $F(1, 21) = 12.63$, $MSe = .022$. Students completed more problems when they spoke to the tutor ($M = .935$, $SD = .112$) than when they typed their responses ($M = .783$, $SD = .156$), $d = 1.12$.

Frequent ASR errors would lead AutoTutor to reject learners' utterances and require them to reiterate their responses, thereby decreasing completion rates. A correlation between the word recognition rate (WRR) and problem completion rates for the spoken condition was not statistically significant, $r(21) = .011$, $p = .961$, so ASR errors did not negatively influence problem completion.

4.2. Learning Gains for Spoken and Typed Conditions

The pretests and posttests were scored for the proportion of questions (out of 24) that participants answered correctly. Proportional learning gains were computed as: [(posttest scores – pretest scores) / (1 – pretest scores)]. Proportional learning gains thus take into account the degree to which participants could improve their score between pretest and posttest.

A repeated measures ANOVA revealed that there were no significant differences in pretest scores across conditions ($p = .537$; see Figure 3). Therefore, any differences in learning across conditions cannot be attributed to differences in prior knowledge. We also correlated pretest scores with WRRs to determine if lower domain knowledge students had increased difficulties with the speech recognition system. However, the results indicated that domain knowledge (pretest scores) and ASR error rates (WRR) were not statistically related, $r(21) = –.161$, $p = .463$.

A repeated measures ANOVA on the proportional learning gains showed a significant difference among conditions, $F(2, 44) = 11.59$, $MSe = .105$. Bonferroni post hoc tests revealed that the significant differences were between the spoken and no tutor conditions ($d = .96$), and between the typed and no tutor conditions ($d = 1.56$). Although there was a .098 ($d = .23$) difference in favor of the typed condition, the difference was not statistically significant.

This lack of a significant difference in learning gains between the two input modalities cannot be simply attributed to our sample size, because our sample size was adequate to detect the 1.0 sigma effect size obtained by other ITSs (Corbett, 2001; Corbett et al., 1999; Graesser et al., 2004; VanLehn et al., 2007) as well as the .8 sigma

FIGURE 3. Learning gains for spoken and typed conditions.

	Spoken		Typed		Control	
Learning Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pretest scores	.321	.192	.261	.176	.299	.172
Posttest scores	.592	.211	.592	.220	.337	.174
Proportional learning gains	.354	.403	.452	.263	.016	.294

mean effect size obtained from previous studies with AutoTutor. It was also large enough to detect the .74 sigma effect reported by Litman et al. (2006; the only available effect size comparing speech and typed input with ITSs). In general, a power analysis (power = .8 and $\alpha = .05$) confirmed that our sample was sufficiently large to detect a medium effect of .54 sigma or higher with a one-tailed paired-samples *t* test (J. Cohen, 1992).

An obvious concern is that the ASR errors might have had a negative impact on learning gains in the spoken condition. However, the correlation between ASR errors and learning gains was not statistically significant, $r(21) = .218$, $p = .319$, which is similar to Litman and colleagues' (2006) results that ASR errors do not appear to have a negative influence on learning gains.

4.3. User Satisfaction for Spoken and Typed Conditions

We performed separate repeated measures ANOVAs on each of the eight questions on the questionnaire to determine whether participants preferred the spoken or typed versions of AutoTutor (see Figure 4). There were significant differences for Items 3, 5, and 7, but not for any of the other questions. Hence, it appears that participants were aware of the ASR errors (Question 3), $F(1, 21) = 7.62$, $MSe = .693$, $d = .48$. Participants felt that the spoken input version of AutoTutor was more difficult to work with than the typed version of AutoTutor (Question 5), $F(2, 44) = 10.01$, $MSe = .370$, $d = .43$. They rated the difficulty of learning computer literacy higher when speaking their responses as opposed to typing (Question 7), $F(2, 44) = 6.44$, $MSe = .599$, $d = .36$.

Although there were no significant differences for Questions 1, 2, 4, and 6, the data reveal that participants preferred the typed interaction. An analysis that investigated whether ASR errors influenced participants' negative evaluations of the spoken-input AutoTutor revealed that there were no significant correlations between WRR and participants' responses to the eight items. This suggests that ASR errors could not explain participants' impressions of the interface. One possibility is that participants feel uncomfortable speaking to a computer and this general discomfort was

FIGURE 4. Learners' evaluations for spoken and typed conditions.

	Spoken		Typed		
Question	*M*	*SD*	*M*	*SD*	*d*
(Q1) I enjoyed the interaction	2.826	1.267	3.304	1.521	.34
(Q2) Interaction comparable to human tutor	2.130	1.058	2.478	1.310	.29
(Q3) Tutor did not understand me*	4.000	1.477	3.304	1.396	.48
(Q4) I was engaged	3.130	1.290	3.565	1.409	.32
(Q5) Tutor was difficult to work with*	4.087	1.203	3.522	1.410	.43
(Q6) I felt like I learned new information	4.217	1.043	4.478	0.898	.27
(Q7) Difficulty of learning computer literacy*	5.000	1.567	4.435	1.532	.36
(Q8) Difficulty of learning new information	4.391	1.672	4.261	1.389	.08

*$p < .05$.

manifested in the ratings. There is yet another possibility that should be mentioned. The learners might have lowered their impression after encountering a very small number of errors in speech recognition. If this is the case, then there is not much of a difference between five errors or 30 errors because both cause an equivalent reduction in learners' impressions. This hypothesis would explain why WRR was not correlated with participants' impressions.

4.4. Conversational Styles for Spoken and Typed Conditions

We mined AutoTutor's log files to obtain a number of features that describe the manner in which participants conversed with AutoTutor. The features were *response verbosity, answer attempts, speech act,* and *word type* (described next). It is important to acknowledge a couple of important points regarding the computing of the various conversational features. To prevent ASR errors from confounding the results, all features were computed from the human transcribed utterances (no errors) rather than the automatically recognized utterances (with errors). For example, if response verbosity was obtained from the automatically recognized texts, then the substantial insertion errors would bias the results. For the same reason we relied on a human rater for the speech act coding rather than AutoTutor's automatic speech act classifier, because we expect the automated system to be affected by ASR errors (investigated in the next section).

Descriptive statistics on the various conversational features are presented in Figure 5. We performed separate repeated measures ANOVAs on each of the features to determine whether there were differences in conversational styles across modalities.

FIGURE 5. Student interaction patterns for spoken and typed conditions.

	Spoken		Typed		
Measure	*M*	*SD*	*M*	*SD*	*d*
Response verbosity					
No. of characters	29.023	17.289	25.077	11.034	.27
Answer attempts					
No. of turns	11.073	1.483	10.476	1.752	.37
Speech act					
Contributions*	0.737	0.152	0.836	0.106	.76
Metacognitive*	0.183	0.128	0.110	0.084	.67
Metacommunicative	0.023	0.027	0.016	0.025	.27
Short responses	0.039	0.032	0.023	0.027	.54
Questions	0.019	0.043	0.018	0.024	.03
Word type					
Self-reference words*	0.050	0.046	0.028	0.026	.59
Social words	0.044	0.021	0.041	0.019	.15
Cognitive words	0.093	0.047	0.076	0.029	.44
Positive emotional words	0.013	0.015	0.014	0.010	.08
Negative emotional words	0.003	0.004	0.003	0.005	.00

*$p < .05$.

Response Verbosity

Response verbosity was measured by the number of characters in the learner's responses. It appears that learners were equally verbose when they typed versus when they spoke their responses ($p = .187$). This finding is inconsistent with previous research by De La Paz and Graham (1997), who found that writers produce text faster and more efficiently while dictating, as compared to writing by hand or typing. We might explain this inconsistency by calling into consideration the differences in the tasks, that is, dictation in their study versus conversational responses to questions in tutoring.

This result is consistent with Litman et al.'s (2006) finding that there were no differences in the average number of words per turn (functionally similar to response verbosity) when spoken human–computer tutorial dialogues were compared to typed tutorial dialogues. However, the lack of a significant difference might be attributed to the small sample sizes in both Litman's and the current experiment.

Answer Attempts

One potential concern of ITSs that support spoken responses is that content coverage will require more time and effort because learners will have to repeat their responses due to ASR errors. Repeated attempts to answer a question after not being understood could heighten learners' frustration. Unresolved frustration could ultimately lead to boredom, which is an emotion that negatively impacts learning (Craig, Graesser, Sullins, & Gholson, 2004). Our results indicate that there were no statistical differences between the spoken and typed conditions for the number of attempts required to answer a question ($p = .102$), although we might not have the requisite statistical power to detect this small effect.

Speech Acts

Two human coders, trained on how to classify speech acts from tutoring dialogues, classified each utterance in the spoken and typed conditions. Occasional disagreements between the coders were resolved in a subsequent coding session. The speech acts were contributions, metacognitive statements, metacommunicative statements, short responses, and questions. Contributions are statements that are directly related to the tutorial domain (e.g., "switch to virtual memory"). Metacognitive statements (e.g., "I don't understand," "I do not know") are indicative of the cognitive state of the student, whereas metacommunicative statements (e.g., "Please repeat," "You are not helping me") refer to the communication-related aspects of the dialogue between the student and the tutor. Finally, short responses include one or two word statements (e.g., "yes" or "no").

An analysis of the proportion of domain specific responses (i.e., contributions) versus other expressions (metacognitive, metacommunicative, short responses, and questions) revealed that learners utilized a reliably higher proportion of other expressions when they spoke their responses compared to when they typed their answers,

$F(1, 21) = 20.48$, $MSe = .006$, $d = .76$. This is an important finding which suggests that learners do respond differently when speaking versus typing.

We compared the proportional usage of each of the four expression types as a function of the spoken versus typed conditions. The results revealed that there were significant differences in the proportion of metacognitive speech acts in favor of the spoken condition, $F(1, 21) = 16.57$, $MSe = .004$, $d = .67$. Therefore, when speaking, students find a greater need to express their cognitive states to the tutor. The most common metacognitive speech act was "I don't know." Other examples are, "Um if I'm understanding your question I think it's yes" and "I have no idea."

There were no significant differences in the proportional usage of metacommunicative expressions, questions, and short responses ($p > .05$), although it is possible that the .54 sigma effect for short responses would have been detected with a larger sample.

Word Usage

We considered five additional features that are sensitive to the type of words used by the participants. These include self-reference words (e.g., "I," "me"), social words (e.g., "you," "us"), cognitive words (e.g., "thinking," "confusing"), positive emotional words (e.g., "happy"), and negative emotional terms (e.g., "sad," "frustrated"). The Linguistic Inquiry and Word Count tool (Pennebaker, Francis, & Booth, 2001) was used to perform the requisite computation.

The results indicated that students used more self-reference words when they spoke to AutoTutor compared to when they typed their responses, $F(1, 21) = 8.29$, $MSe = .001$, $d = .59$. The usage of social, positive, and negative emotional terms was indistinguishable across conditions, although there was a trend in the spoken direction for cognitive terms.

Relationship Between Conversational Styles and ASR Errors

We conducted a follow-up analysis to investigate whether it was the ASR errors that caused students to increase their usage of metacognitive speech acts and self-reference words when speaking their responses to AutoTutor. Correlational analyses did not yield any significant relationships between word recognition rate and these features. There was also no relationship between usage of metacognitive expressions, self-reference words, and pretest scores, so prior knowledge cannot explain the increased use of these features in spoken dialogues.

However, these features were significantly related to the number of attempts to answer the main question, $r(21) = .595$ for metacognitive expressions and $r(21) = .696$ for self-reference terms. These correlations suggest that students use these expressions when they are having trouble with the material and make repeated attempts to answer the main question. It is informative to note that this relationship applies only when learners have a spoken dialogue with AutoTutor; these variables were not related for typed dialogues. This difference suggests that switching the modality from spoken

to typed causes students to adapt their interaction styles when they are having difficulties in mastering the content.

5. RESULTS: EFFECT OF ASR ERRORS ON THE TUTOR (GOALS 5–6)

AutoTutor's SAC and Assessments module are expected to be affected by ASR errors as they are natural language processors. We predict that ASR errors would have a pronounced impact on the SAC module, which performs a syntactic-semantic analysis of learners' utterances. ASR errors are expected to have a comparatively smaller impact on the Assessments module, which is a semantic- and statistical-based natural language processer. This possibility was tested by analyzing the impact of speech recognition errors on each of these modules (Goal 5: Comparatively deep versus shallow NLP).

5.1. Speech Act Classification

The human-coded speech acts (coding is described previously) were compared to the automatically classified speech acts. A SAC error occurred when there was a discrepancy between human-coded speech acts and the automatically classified speech acts. We restricted our analysis to a binary classification decision, that is, was the utterance a *contribution* or some *other* expression (e.g., question, short response, metacognitive, metacommunicative), because this is the first decision that AutoTutor makes when interpreting a learner's response.

A repeated measures ANOVA was used to determine if there were significant differences in error rates (number of incorrectly classified utterances divided by the total number of utterances) associated with the ASR recognized responses (with speech errors) versus the human-transcribed responses (no errors). The results indicated that there were significantly more errors with the ASR responses (M = .123, SD = .114) than the human-transcribed responses (M = .037, SD = .055), $F(1, 22)$ = 16.711, MSe = .005, d = .96. Hence, ASR errors had a negative impact on the accuracy of AutoTutor's SAC system.

This finding can be easily understood once some of the details of the internal mechanisms of the SAC system are explained. AutoTutor's SAC performs a two-step analysis of learners' responses. The classifier first performs part-of-speech tagging on the utterance and then relies on a cascade of finite state transducers that operate on the tagged text (Olney et al., 2003). The transducers rely heavily on the output of the part-of-speech tagger, which in turn relies on the syntactic integrity of the input utterance. However, the ASR system performs poorly when syntax is considered in the performance evaluation (word recognition rate = .54). Therefore, incorrectly recognized speech will have a negative impact on the tagger, which adversely affects the finite state transducers.

These results suggest a limitation of the spoken-input version of AutoTutor. The accuracy of the SAC system is adversely affected by errors in automatically recognized spoken text, $r(21) = -.518$ for the correlation between WRR and SAC errors. These classification errors have quite an impact on the overall interaction experience. More specifically, although SAC errors were not significantly correlated with learning gains, $r(21) = -.158$, $p = .470$, they were negatively correlated with response verbosity, $r(21) = -.460$, and positively correlated with the number of answer attempts, $r(21) = .575$.

SAC errors were also related to the manner in which participants expressed their responses. SAC errors were positively correlated with proportional usage of meta-cognitive speech acts, $r(21) = .422$, self-reference words, $r(21) = .797$, and cognitive words, $r(21) = .727$. The robustness of these correlations indicates that participants substantially adapt their response styles and word usage in the face of ASR-related SAC errors. Therefore, if it is possible to design a comparatively accurate SAC module that uses statistical shallow NLU instead of brittle symbolic NLU, then such a module can be expected to be more resilient to ASR errors and improve learner satisfaction.

5.2. Meaning Assessment Module

AutoTutor's Assessment module has two primary functions. First, it evaluates the conceptual quality of the student's response to determine the appropriate level of feedback and to select the next dialogue move. Second, it uses its assessment of the student's response to update its model of student knowledge across the session. This phase is critical for ensuring that AutoTutor is dynamically responsive to each individual student.

Once the Assessment module receives a student's contribution, it executes two separate computational linguistic analyses. First, each contribution is compared with all of the expectations (ideal answers) for a problem using Latent Semantic Analysis (LSA; Landauer et al., 2007). LSA is a statistical technique that measures the conceptual similarity of two text sources. In this similarity matching algorithm, a vector representing the semantic content of the contribution is created and compared with one or more expectations. The cosine between the two vectors is calculated to produce a match similarity score from 0 to 1 (negative cosines are converted to 0; Graesser et al., 2007).

Once the LSA analysis is completed, an inverse word frequency weighted overlap (IWFWO) analysis is performed. The IWFWO algorithm is a word-matching algorithm in which each word is weighted on a scale from 0.0 to 1.0, relative to its inverse frequency in the English language using the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995). As a consequence, higher frequency words such as closed-class function words (e.g., *and, but, a, the,* etc.) have comparatively low weights and little effect on the IWFWO match score. Lower frequency words (e.g., *RAM, system, speed*, etc.) have higher weights and exert more influence on the IWFWO match score. Similar to LSA, the IWFWO algorithm also generates a semantic match similarity score between 0 and 1 for each expectation.

Both LSA and IWFWO have strengths and weaknesses which are extensively discussed in previous publications (Chipman, 2008; Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007). Hence, AutoTutor uses a hybrid match score ($HYBRID = .33 \times LSA + .67 \times IWFWO$) to leverage the benefits of both systems. By combining the two similarity scores into a semantic match score, the benefits of both models can be realized while minimizing the effects of their flaws.

This hybrid measure is computed separately for each expectation of the current problem. It is referred to as the *semantic match* between the student contribution and any individual expectation. The semantic match is used to determine the quality of the student contribution, to generate AutoTutor's feedback (positive, neutral, negative), and to formulate tutor moves that advance the conversation. The set of semantic match scores for any particular contribution represents AutoTutor's model of student knowledge on each student turn. The student model over multiple turns is based on these semantic match scores.

The aforementioned specification for a single student–tutor turn pair does not tell the whole story, however. If the student model was completely rebuilt on each student turn, the semantic matches would wildly vary, representing the vicissitudes of the student's ability to provide content in response to different tutor dialogue moves. To maintain continuity of the student model between turns, the semantic match is calculated using both the current student response alone (local assessment) and the current response concatenated with all previous student responses for the current problem (global assessment). This global assessment represents AutoTutor's model of student knowledge across turns.

We investigated whether ASR errors negatively influenced AutoTutor's assessments of the learners' responses in their turns. The semantic match between the expectations and the automatically recognized speech (SM_{ASR}) was compared to the semantic match between the expectations and the human transcribed speech ($SM_{TRANSCRIPT}$). If AutoTutor's Assessment module is not influenced by ASR errors, then, $SM_{ASR} = SM_{TRANSCRIPT}$. On the other hand, ASR errors would have an adverse impact on AutoTutor's Assessment module if $SM_{ASR} \neq SM_{TRANSCRIPT}$.

The analyses proceeded in the following manner. First, global semantic match scores for LSA, IWFWO, and Hybrid were computed for each of the spoken contributions by computing the semantic match between the ASR text and each expectation. Because AutoTutor compares students' contributions to multiple expectations, we averaged semantic match scores across the expectations. Second, semantic match scores were computed for each of the participants by averaging the scores obtained from their individual contributions. This procedure yielded three sets of SM_{ASR} scores for LSA, IWFWO, and Hybrid with 23 scores in each set. The procedure was repeated for the human transcribed responses yielding three sets of $SM_{TRANSCRIPT}$ scores. Three paired-samples t tests were used to statistically compare SM_{ASR} and $SM_{TRANSCRIPT}$ scores obtained with each algorithm.

The results indicated that the three metrics of semantic evaluation (i.e., global hybrid match, LSA, and IWFWO scores) showed no significant differences between the students' utterances with ASR errors (SM_{ASR}) and the human transcribed utterances

($SM_{TRANSCRIPT}$), $p > .05$ (see Figure 6). Therefore, it appears that AutoTutor's assessments of student knowledge across turns was not affected by ASR errors. The statistical algorithms are impressively robust as indicated by the correlations between SM_{ASR} and $SM_{TRANSCRIPT}$ scores.

5.3. Word Error Rate Simulations to Generalize Findings

Our results so far indicate that AutoTutor's use of shallow statistical NLU techniques proved to be resilient to ASR errors. Comparisons between semantic matches generated by ASR text and human transcribed text yielded an approximately null effect ($d = .02$) for the hybrid algorithm. However, this conclusion should be interpreted with a modicum of caution because they only apply to one profile of ASR errors. It is plausible that the semantic matching algorithms might produce different results with different speech recognizers. Therefore, for our findings to be applied to other systems, we conducted a simulation to model the behavior of the semantic matching algorithms when confronted with different error profiles (Goal 6: ASR error simulation).

The simulation proceeded by using the human transcribed utterances (no errors) to construct surrogate data sets with simulated ASR errors. The WER for AutoTutor was .45 with the distribution of errors being: substitution errors = 52%, deletion errors = 9%, and insertion errors = 39%. Simulated word error rates ranged from 0.1 to 0.8 with increments of 0.1, with the distribution of errors preserved.

Substitution errors were simulated by randomly selecting a word in the human-transcribed utterance (error = 0) and replacing it with a word that was randomly selected from the ASR's dictionary. The random word was usually a content word (i.e., *RAM, hardware*), but it could sometimes be a common word (e.g., *and, but, because*). For deletion errors, a randomly selected word from the student's utterance was deleted. We simulated insertion errors by randomly selecting a word from the dictionary and inserting it at a random position in the utterance.

This error simulation process was repeated for each participant's data, yielding 207 data sets (23 participants × 9 error rates). Each data set was submitted to AutoTutor's Assessment module, and global semantic matches between the students' utterance and the expected answers were computed by LSA, IWFWO, and the hybrid matching algorithm.

FIGURE 6. Tutor assessments of student knowledge across turns.

	SM_{ASR}		$SM_{TRANSCRIPT}$			
Measure	*M*	*SD*	*M*	*SD*	*d*	*r*
Hybrid	0.195	0.104	0.193	0.099	.02	.947
LSA	0.165	0.067	0.173	0.062	.12	.991
IWFWO	0.210	0.127	0.203	0.122	.06	.989

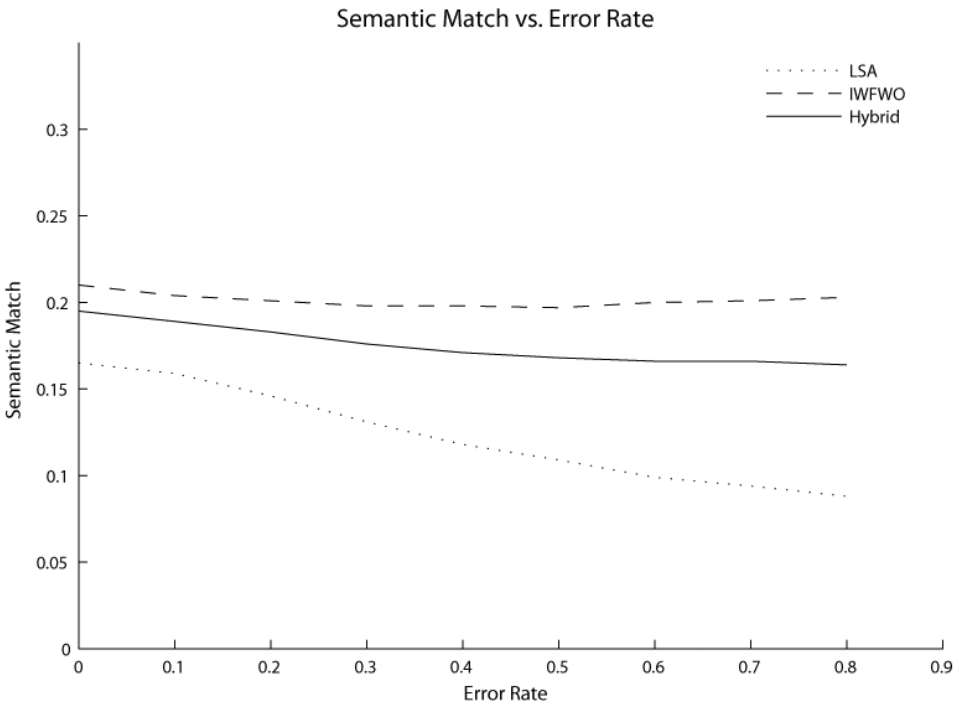
Note. SM_{ASR} = automatically recognized speech; $SM_{TRANSCRIPT}$ = human transcribed speech; LSA = Latent Semantic Analysis; IWFWO = inverse word frequency weighted overlap.

We performed a 9 × 3 repeated measures ANOVA with the semantic match score as the dependent variable and *error rate* and *match type* as within subject factors. Error rate had nine levels ranging from 0.0 (no error) to 0.8 (substantial errors). Match type had three levels for LSA, IWFWO, and HYBRID.

Not surprisingly, the main effect for error rate was statistically significant, $F(8, 176) = 38.06$, $MSe = .0004$, $p < .001$, partial $\eta^2 = .634$. Hence, the semantic match between participants' responses and expectations (good answers) decreased as ASR errors increased. The main effect for match type was also significant, $F(2, 44) = 13.67$, $MSe = .00006$, $p < .001$, partial $\eta^2 = .383$, indicating that there were differences in semantic match scores produced by the three algorithms.

The most interesting finding was the Error Rate × Match Type interaction, which was statistically significant, $F(16, 352) = 43.66$, $MSe = .00009$, $p < .001$, partial $\eta^2 = .665$. This interaction is presented in Figure 7. It appears that semantic match scores provided by LSA were negatively impacted by ASR errors (sharp linear decrease with slope = –.103 and $R^2 = .980$). IWFWO scores, on the other hand, appear to be resilient to ASR errors (approximately flat line with slope = –.007 and $R^2 = .208$). Semantic matches produced by the hybrid model, that combined LSA and IWFWO, decreased at a higher rate than IWFWO but slower than LSA (slow linear decrease with slope = –.039 and $R^2 = .913$). Taken together, these results suggest that LSA deteriorates rap-

FIGURE 7. ASR Error Rate × Semantic Matching Algorithm interaction. *Note*. LSA = Latent Semantic Analysis; IWFWO = inverse word frequency weighted overlap.



idly in face of ASR errors, but IWFWO and the Hybrid algorithm are more robust to these errors.

6. GENERAL DISCUSSION

This research was motivated by the assumption that the communication bandwidth between a human student and a computer tutor may be substantially enhanced through spoken student responses. This assumption was investigated through a broad set of analyses of both the human and the computer. Goals 1 to 4 assessed the influence of ASR-based human–computer spoken tutorial dialogues on content coverage, learning gains, learner satisfaction, and learners' conversational styles. Goals 5 and 6 addressed how ASR errors impact the computer tutor with respect to the SAC module and the Assessment module's three semantic algorithms that assess the student's contributions. The subsequent discussion discusses six major findings, each aligned with the goals proposed in the Introduction.

The first finding is that students completed more problems in the spoken condition, ostensibly due to the efficiency of speech production (De La Paz & Graham, 1997). This enhanced content coverage, which is equivalent to shorter task completion times, seems to be a highly desirable and generalizable feature of spoken interactions (Allen, Miller, Ringger, & Sikorski, 1996; Damianos, Loehr, Burke, Hansen, & Viszmeg, 2003; Harris & Biermann, 2002; Litman et al., 2006). But merely completing more problems surprisingly did not translate into enhanced learning gains, $r(21) = -.129$, $p = .557$. We suspect that this result might be a function of the short training (35 min) of the tutorial interventions in the current study. Perhaps the merits of enhanced content coverage could be realized over longer training sessions of 2 hr or longer. This is an empirical question that warrants further research.

Our second important finding was that changing the input modality from typed to spoken did not yield increased learning gains. The current analysis attempted to replicate the results reported by Litman et al. (2006) in their comparison of spoken and typed *computer tutorial* interactions. Litman et al. investigated physics whereas we investigated computer literacy, but we had expected similar results to emerge. Aside from attempting to replicate the Litman results, our study addressed a potential confound with Litman's experiment. The ITSPOKE tutor in Litman's study presented a dialogue history of the student's spoken responses, so the students were aware of ASR errors. It is conceivable that the students might have lost confidence in the computer tutor due to its inability to fully comprehend their responses. This threat to communication fidelity or the lack of confidence might have negatively influenced their ability to learn from the computer tutor. The version of AutoTutor in the current study did not give students access to a dialogue history, so the possibility of speech recognition errors was not explicitly manifested. Although concealing ASR errors from the students did not change the pattern of the results, at the very least, one confound in the previous study was ruled out in the current experiment.

It is important to acknowledge some limitations with our methodology that might have contributed to the lack of significant difference between conditions. One limitation is our small sample size of 23 participants. Although this sample had adequate statistical power to a medium effect size of .54 sigma or higher, the learning gains data (Figure 3) suggest that a smaller effect ($d = .29$ in favor of the typed condition) occurred in our sample of college students. Our experiment did not have sufficient power to detect this small effect. Another limitation pertains to our use of a pretest, which could have resulted in a testing effect (Carrier & Pashler, 1992; Roediger & Karpicke, 2006). It might also be the case that the four-alternative multiple-choice tests that were used to compute learning gains were not sufficiently sensitive to detect subtle effects between conditions. Hence, a replication with a larger sample, posttest-only design, and open-ended knowledge tests is warranted to alleviate these limitations.

Aside from these qualifications, it is important to note that the small effect in favor of typed dialogues suggests that there are some unique advantages of typing. One advantage of typed input is that students can evaluate their responses, remove errors or misspellings, and revise their contributions. The additional time to reflect on their composition might yield superior learning gains compared to speech-based ITSs in which students have to generate their responses in real time (Quinlan, 2004). Yet another advantage of typed input is that the students have time to plan their contributions, pause, and think, whereas spoken input normally is accompanied with expectations to provide continuous input. Planning is associated with deeper cognitive processing and presumably increased learning gains.

The third important finding is that learners seemed to prefer typing their responses over speaking. Although this preference was not linked to learning gains over a short 35-min tutorial session, it is likely that it might have a more profound impact on learning when tutorial sessions are expanded to a week, a semester, or even a year. It is possible that the dislike of spoken utterances was a reflection of poor speech recognition performance. Participants may have given the spoken-input method low ratings due to frustration or disappointment with the quality of the speech recognition system, although survey outcomes did not correlate with word recognition rates. An alternative interpretation would be to attribute learners' preference for typed interventions to the novelty of the spoken interaction. Students are not used to speaking aloud to computers, so it is quite plausible that enhanced training periods will provide students with the necessary experience to overcome their initial awkwardness and negative perceptions of the spoken-input tutor.

In summary, two possible factors might explain learners' preferences of typing versus speaking. The first is system failure via ASR and SAC errors. The second is the learners' unfamiliarity and general dislike of spoken interfaces. We conducted a follow-up Wizard-of-Oz-type experiment to determine which of these factors played a role in students' perceptions of the spoken-input AutoTutor. The follow-up experiment was identical to the current experiment in all respects but with one important exception. Instead of the commercially available recognizer used in the current experiment (WRR = .54), an experimenter in an adjacent room manually performed the speech recognition in real-time, yielding highly accurate speech recognition (WRR =

.90). The results indicated that in contrast to the current experiment, where learners preferred typing over speaking, learners in the follow-up experiment showed no preference for input modality ($N = 24$, power = .67). The implications of these results are that it is not the learner's preconceived notions and general dislike of spoken interfaces but rather system errors that negatively influence their perceptions of spoken tutorial dialogues. Therefore, reengineering AutoTutor's SAC system to be more resilient to ASR errors is expected to improve learner satisfaction.

The fourth finding was that our results partially substantiated predictions of a student-oriented extension of the social agency theory (Louwerse, Graesser, Lu, & Mitchell, 2005; Mayer et al., 2003). It appears that participants used more metacognitive and self-reference terms when speaking to the tutor, as would be predicted by this theory. Although it is tempting to speculate that participants interacted differently when speaking rather than typing for socially motivated reasons, it is possible that the use of these expressions might be related to increased answer attempts when the AutoTutor's SAC system is faulty. More research is needed to differentiate these two possible explanations.

Our fifth research finding was that the SAC module appeared to be quite brittle in the face of imperfect input. In contrast, the Assessment module was comparatively impervious to these errors. Therefore, SAC accuracy of AutoTutor might be improved by utilizing more semantic and statistical rather than syntactic- and symbol-based NLU mechanisms.

In addition to the limitations of the NLU mechanisms used by the SAC, differences in learners' communication styles when speaking versus typing might also be related to the lower performance. AutoTutor's SAC module was trained on utterances obtained from several previous experiments where learners *typed* their responses. It is reasonable to assume that the content of participants' verbal expressions differed as the modality changed from typed to spoken input (Litman et al., 2006). Therefore, the existing SAC system may have been unable to accommodate the modified conversational styles. This limitation can be corrected by retraining the SAC module to detect spoken conversational patterns.

The sixth finding of this research was that the IWFWO and LSA semantic match algorithms were differentially affected by ASR errors. LSA faltered when confronted with an increase in ASR errors, whereas the IWFWO and the hybrid algorithms were fairly resilient to ASR errors. These results suggest that the most defensible solution is a hybrid algorithm that uses both of these mechanisms, following a soft-constraint satisfaction (SCS) approach. According to SCS models, the performance of an intelligent system should not rely on the integrity of any one level or module but rather should reflect the confluence of several levels or modules that are statistically combined. For example, natural language understanding involves a multilevel computational analysis including phonemes, morphemes, words, syntax, sentence semantics, discourse, pragmatics, world knowledge, and genre. According to an SCS model, when one level fails, the other levels fill in. When there is an ambiguity at one level, the other levels resolve the ambiguity. For example, context can be recruited to resolve ambiguity in word meaning (Pickering & Garrod, 2004; Waltz & Pollack, 1985). The hybrid algorithm can

be considered to be a within-level SCS method because multiple algorithms are recruited to resolve ambiguity at the same level (i.e., semantic level) rather than across levels (e.g., phonemes, morphemes, words, syntax, etc.).

The fact that these findings were obtained by simulated WER increases their generalizability to other computer interfaces that support spoken dialogues. These simulation results, however, should be interpreted with a modicum of caution because a naive algorithm that randomly inserted, deleted, and substituted words was used to simulate ASR errors. A more realistic algorithm would simulate errors that reflected the acoustic confusability of words in a context specific fashion (Pietquin & Dutoit, 2006; Schatzmann, Thomson, & Young, 2007). Such an algorithm is preferred over the naive algorithm because a word in the utterance is replaced with a similar-sounding word instead of a randomly selected word. The current analyses opted for the naive algorithm by virtue of its simplicity, but more realistic error simulation algorithms will be considered in the future. To this point, a recent paper compared a naive error simulation algorithm to a more realistic algorithm that used weighted finite state transducers over a range of word error rates (Stuttle, Williams, & Young, 2004). The results revealed that the naive algorithm degraded more rapidly as simulated WER increased, so the naive algorithm represents a lower bound on performance in light of ASR errors.

7. CONCLUSIONS

A synthesis of the results of the current study, related research, and our theoretical perspectives on pedagogy and tutorial dialogue paints a mixed picture of the merits of spoken tutorial dialogues. The speech facilitation hypothesis was supported neither in the current experiment with AutoTutor nor in the previous experiment with ITSPOKE (Litman et al., 2006). ASR errors cannot be linked to the lack of significant learning gains in these experiments because these errors were not correlated with learning. Furthermore, learning gains in the spoken and typed condition were equivalent in the Wizard-of-Oz study (described earlier) where there were relatively few ASR errors.

The lack of a modality effect in these three experiments with computer tutors suggests a *modality equivalence hypothesis*. This hypothesis claims that the effects of the input media (spoken or typed) on learning gains are either subtle or nonexistent. Simply put, the *content* is more important than the *medium* of communication (Graesser et al., 2003). Or put more dramatically, as a counterpoint to Marshall McLuhan (1964): "The medium is not the message—The message is the message."

The modality equivalence hypothesis questions the utility of migrating to spoken input. This is because compared to spoken dialogues, typed dialogues are computationally easier to implement and equally effective. But before we accept this conclusion too cavalierly, there is one unaccounted effect that needs to be addressed. Recall that Litman et al.'s (2006) experiment with a *human* tutor indicated that spoken input resulted in significantly more learning when compared to typed input.

One possible explanation for Litman's effect with the human tutor is that interruptions and overlapping speech were allowed to occur in the spoken condition but not the typed condition (Litman et al., 2006). It is quite possible that the enhanced communication bandwidth offered by spontaneous speech might have resulted in its heightened effectiveness. For example, the tutor could infer that the learner was experiencing uncertainty by monitoring speech disfluencies, hesitations, and other perturbations in the learner's response (Forbes-Riley & Litman, 2009). The tutor could respond to the learner's uncertainty by offering a hint or a simplified problem. This type of just-in-time dynamic response was not permitted in the typed condition where strict turn-taking was enforced.

Aside from this potential confound with Litman's human–human tutoring study, the fact that the speech-facilitation hypothesis was not replicated in the three experiments with the computer tutors indicates that there is something about spoken human tutoring that is missed by the computer tutors. In our view, the missing link lies in the information that can be extracted from the paralinguistic features of the student's speech. Monitoring these nonverbal speech cues provides a rich trace of information related to the social dynamics, attitudes, urgency, and emotions of the learner (Batliner, Fischer, Huber, Spilker, & Noth, 2003; Bosch, 2003; Litman & Forbes-Riley, 2004; Scherer, 2003; Shafran, Riley, & Mohri, 2003).

It is no surprise that studies with computer tutors that ignore these information channels show equivalent trends between spoken and typed dialogues (like the current study and Litman's 2006 study). Hence, contradictory to theories that support the speech facilitation hypothesis, the cause of the effect with human tutors lies not in the student's speech but in the tutor's interpretation of the speech signal above and beyond its verbal content. Hence, developing systems that monitor the paralinguistic features of speech might be the key to bridging the performance gap for spoken dialogues with human tutors versus artificial tutoring systems.

In particular, information extracted from the paralinguistic features of speech could be leveraged in at least two significant ways. First, smooth conversation requires dialogue management. There need to be discourse markers and other cues that guide the student in the exchange (Core, Moore, & Zinn, 2003; Freedman, 1996; Moore, 1995). A collaborative exchange between a computer tutor and the learner requires a mutual understanding of the turn-taking process. In human-to-human conversations, speakers signal to listeners that they are relinquishing the floor and that it is the listener's turn to say something (Clark, 1996; Sacks, Schegloff, & Jefferson, 1978). However, typed human-to-computer conversations lack many of the subtle signals inherent to human conversations. When conversational agents lack turn-taking signals, the learner does not know when or if the learner is supposed to respond and is sometimes confused when the tutor generates particular dialogue moves. These problems can be minimized by detecting turn-boundaries by monitoring the paralinguistic features of speech akin to human–human communication.

Second, increased interest in the role of affect in learning and tutoring has led to recent work on affect-sensitive ITSs (D'Mello, Craig, Fike, & Graesser, 2009; D'Mello, Picard, & Graesser, 2007; Picard, 1997). Systems that can utilize the prosodic and

acoustic characteristics of students' utterances in their decision-making processes will be better equipped to deal with bored, frustrated, and confused students than text-based systems that do not have access to these features. If the learner is frustrated, for example, the tutor can give hints to advance the learner in constructing knowledge or can make supportive empathetic comments to enhance motivation. If the learner is bored, the tutor needs to present more engaging or challenging problems for the learner to work on. The tutor would probably want to lay low and stay out of the learner's way when the learner is deeply engaged in learning. Whether such systems positively influence learning and engagement as well as the best human tutors awaits further development and empirical testing.

NOTES

Acknowledgments. We thank our research colleagues in the Emotive Computing Group and the Tutoring Research Group at the University of Memphis (http://emotion.autotutor.org). Special thanks to Jeremiah Sullins and O'meed Entezari for their valuable contributions to this study.

Support. This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, and HCC 0834847). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of NSF.

Authors' Present Addresses. Sidney D'Mello, 202 Psychology Building, University of Memphis, Memphis, TN 38152. E-mail: sdmello@memphis.edu. Art Graesser, 202 Psychology, University of Memphis, Memphis, TN 38152. E-mail: graesser@memphis.edu. Brandon King, Center for Mind and Brain, University of California, Davis, 267 Cousteau Place, Davis, CA 95618. E-mail: bgking@ucdavis.edu

HCI Editorial Record. Received September 8, 2008. Revision received June 17, 2009. Accepted by John Anderson. Final manuscript received December 11, 2009. — *Editor*

REFERENCES

Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147–179.

Allen, J. (1995). *Natural language understanding* (2nd ed.). Redwood City, CA: Benjamin/Cummings.

Allen, J., Miller, B., Ringger, E., & Sikorski, T. (1996). A robust system for natural spoken dialogue. In A. Joshi & M. Palmer (Eds.), *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 62–70). San Francisco, CA: Morgan Kaufmann.

Anderson, J., Corbett, A., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4,* 167–207.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD–ROM).* Philadelphia, PA: University of Pennsylvania.

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

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Bosch, L. (2003). Emotions, speech and the ASR framework. *Speech Communication, 40,* 213–225.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35–53). Norwood, NJ: Ablex.

Chi, M., Roy, M., & Hausmann, R. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–341.

Chi, M., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22,* 363–387.

Chi, M., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–533.

Chipman, P. (2008). *An analysis and optimization of AutoTutor's student model* (Unpublished master's thesis). University of Memphis, Memphis, Tennessee.

Clark, H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, P., Kulik, J., & Kulik, C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. *Proceedings of the Eighth International Conference on User Modeling.* Berlin, Germany: Springer.

Corbett, A. (2002). Cognitive tutor algebra I: Adaptive student modeling in widespread classroom use. *Proceedings of the Technology and Assessment: Thinking Ahead Workshop.* Washington, DC: National Academies Press.

Corbett, A., Anderson, J., Graesser, A., Koedinger, K., & VanLehn, K. (1999). Third generation computer tutors: Learn from or ignore human tutors? *Proceedings of the CHI Conference on Human Factors in Computing Systems.* New York, NY: ACM.

Core, M., Moore, J., & Zinn, C. (2003). The role of initiative in tutorial dialogue. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics.* Philadelphia, PA: Linguistic Data Consortium.

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

Dalgarno, B. (2001). Interpretations of constructivism and consequences for computer assisted learning. *British Journal of Educational Technology, 32,* 183–194.

Damianos, L., Loehr, D., Burke, C., Hansen, S., & Viszmeg, M. (2003). The MSIIA experiment: Using speech to enhance human performance on a cognitive task. *International Journal of Speech Technology, 6,* 133–144.

De La Paz, S., & Graham, S. (1997). Effects of dictation and advanced planning instruction on the composing of students with writing and learning problems. *Journal of Educational Psychology, 89,* 203–222.

D'Mello, S., Craig, S., Fike, K., & Graesser, A. (2009). Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In J. Jacko (Ed.), *Human–computer interaction. Ambient, ubiquitous and intelligent interaction* (pp. 595–604). Berlin, Germany: Springer.

D'Mello, S., Picard, R., & Graesser, A. (2007). Towards an affect-sensitive AutoTutor. *Intelligent Systems, IEEE, 22*(4), 53–61.

Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. In V. Dimitrova, R. Mizoguchi, & B. Du Boulay (Eds.), *Proceedings of the 14th Inter-*

national Conference on Artificial Intelligence in Education (pp. 33–40). Amsterdam, the Netherlands: IOS Press.

Freedman, R. (1996). *Interaction of discourse planning, instructional planning, and dialogue management in an interactive tutoring system* (Unpublished doctoral dissertation). Northwestern University, Evanston, Illinois.

Gertner, A., & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 133–142). Berlin, Germany: Springer.

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

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., & Louweres, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36*, 180–193.

Graesser, A., Moreno, K., Marineau, J., Adcock, A., Olney, A., & Person, N. (2003). AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head? . In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 47–54). Amsterdam, the Netherlands: IOS Press.

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What is a good question? In M. McKeown & G. Kucan (Eds.), *Bringing reading research to life* (pp. 112–141). New York, NY: Guilford.

Graesser, A., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

Graesser, A., Person, N., Lu, Z., Jeon, M., & McDaniel, B. (2005). Learning while holding a conversation with a computer. In L. PytlikZillig, M. Bodvarsson, & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 143–167). Greenwich, CT: Information Age.

Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9,* 495–522.

Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8,* 129–147.

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication, 49*, 861–873.

Harris, S., & Biermann, A. (2002). Mouse selection versus voice selection of menu items. *International Journal of Speech Technology, 5,* 398–402.

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A., & McNamara, D. (2007). Strengths, limitations, and extensions of LSA. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 401–426). Mahwah, NJ: Erlbaum.

Johnson, W., & Valente, L. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. *Proceedings of the 20th National Conference on Innovative Applications of Artificial Intelligence*. New York, NY: AAAI Press.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kato, K., Nanjo, H., & Kawahara, T. (2000). Automatic transcription of lecture speech using topic-independent language modeling. *Proceedings of the 6th International Conference on Spoken Language Processing*. International Speech Communication Association (ISCA).

Koedinger, K., Anderson, J., Hadley, W., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8,* 30–43.

Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Leeuwis, E., Federico, M., & Cettolo, M. (2003). Language modeling and transcription of the TED corpus lectures. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* New York, NY: IEEE.

Lesgold, A., Lajoie, S., Bunzo, M., & Eggan, G. (1992). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. H. Larkin & R. W. Chabay (Eds.), *Computer-assisted instruction and intelligent tutoring systems* (pp. 201–238). Hillsdale, NJ: Erlbaum.

Levinson, S. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Lickley, R. (1994). *Detecting disfluency in spontaneous speech* (Unpublished doctoral dissertation). University of Edinburgh, Edinburgh, Scotland.

Litman, D., & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.* Stroudsburg, PA: Association for Computational Linguistics.

Litman, D., Rose, C., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education, 16,* 145–170.

Litman, D., & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. *Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics.* Stroudsburg, PA: Association for Computational Linguistics.

Louwerse, M., Graesser, A., Lu, S., & Mitchell, H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19,* 693–704.

Mayer, R. (Ed.). (2005). *The Cambridge handbook of multimedia learning.* New York, NY: Cambridge University Press.

Mayer, R., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology, 96,* 389–395.

Mayer, R., Sobko, K., & Mautone, P. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95,* 419–425.

McLuhan, M. (1964). *Understanding media: The extensions of man.* London, UK: Routledge.

Moore, J. (1995). *Participating in explanatory dialogues.* Cambridge, MA: MIT Press.

Moshman, D. (1982). Exogenous, endogenous, and dialectical constructivism. *Developmental Review, 2,* 371–384.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169–234). Cambridge, MA: MIT Press.

Munteanu, C., Penn, G., Baecker, R., & Zhang, Y. (2006). Automatic speech recognition for webcasts: How good is good enough and what to do when it isn't. *Proceedings of the 8th International Conference on Multimodal interfaces.* New York, NY: ACM.

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). Utterance classification in AutoTutor. *Proceedings of the Human Language Technology - North American Chapter of the Association for Computational Linguistics Conference.* Stroudsburg, PA: Association for Computational Linguistics.

Paas, F., van Merrienboer, J., & Adam, J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*(1), 419–430.

Pellom, B., & Hacioglu, K. (2003). Recent improvements in the CU Sonic ASR system for noisy speech: the SPINE task. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,*4–7. IEEE.

Pennebaker, J., Francis, M., & Booth, R. (2001). *Linguistic inquiry and word count (LIWC): A computerized text analysis program.* Mahwah, NJ: Erlbaum.

Person, N., & Graesser, A. (2002). Human or computer? AutoTutor, in a bystander turing test. *Proceedings of the 6th International Conference on Intelligent Tutoring Systems.* Berlin, Germany: Springer.

Picard, R. (1997). *Affective computing.* Cambridge, MA: MIT Press.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169–225.

Pietquin, O., & Dutoit, T. (2006). A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech, and Language Processing, 14,* 589–599.

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E. O., & Peters, S. (2004). Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. *Proceedings of Seventh International Conference on Intelligent Tutoring Systems.* Berlin, Germany: Springer.

Psotka, J., Massey, D., & Mutter, S. (1988). *Intelligent tutoring systems: Lessons learned*: Hillsdale, NJ: Erlbaum.

Quinlan, T. (2004). Speech recognition technology and students with writing difficulties: Improving fluency. *Journal of Educational Psychology, 96,* 337–346.

Roediger, H., & Karpicke, J. (2006). Test-enhanced learning - Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255.

Rogina, I., & Schaaf, T. (2002). Lecture and presentation tracking in an intelligent meeting room. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces.* New York, NY: IEEE.

Rus, V., & Graesser, A. (2007). Lexico-syntactic subsumption for textual entailment. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Recent advances in natural language processing IV: Selected papers from RANLP 2005* (pp. 187–196). Amsterdam, the Netherlands: John Benjamins.

Sacks, H., Schegloff, E., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction.* New York, NY: Academic Press. pp. 7–55.

Schatzmann, J., Thomson, B., & Young, S. (2007). Error simulation for training statistical dialogue systems. *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding.* New York, NY: IEEE.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40,* 227–256.

Schultz, K., Bratt, E., Clark, B., Peters, S., Pon-Barry, H., & Treeratpituk, P. (2003). A scalable, reusable spoke conversational tutor: Scot. *Proceedings of the Workshop on Tutorial Dialogue Systems: With a View toward the Classroom. In conjunction with the 11th International Conference on Artificial Intelligence in Education.*

Shafran, I., Riley, M., & Mohri, M. (2003). Voice signatures. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.* New York, NY: IEEE.

Shah, F., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes, 33,* 23–52.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies* (Unpublished doctoral dissertation). Stanford University, Palo Alto, California.

Sleeman, D., & Brown, J. (Eds.). (1982). *Intelligent tutoring systems.* New York, NY: Academic Press.

Storey, J., Kopp, K., Wiemer, K., Chipman, P., & Graesser, A. (in press). Critical thinking tutor: Using AutoTutor to teach scientific critical thinking skills. *Behavioral Research Methods.*

Stuttle, M., Williams, J., & Young, S. (2004). A framework for dialog systems data collection using a simulated ASR channel. *Proceedings of the International Conference on Spoken Language Processing.* International Speech Communication Association (ISCA).

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language, 58*(1), 1–21.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16,* 227–265.

VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3–62.

Waltz, D., & Pollack, J. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9*(1), 51–74.

Whittaker, S. (2003). Theories and methods in mediated communication. In A. Graesser, M. Gernsbacher, & S. Goldman (Eds.), *The handbook of discourse processes* (pp. 243–286). Mahwah, NJ: Erlbaum.

Woolf, B. (2009). *Building intelligent interactive tutors.* Burlington, MA: Morgan Kaufmann.

Zolnay, A., Kocharov, D., Schluter, R., & Ney, H. (2007). Using multiple acoustic feature sets for speech recognition. *Speech Communication, 49*, 514–525.