Does It Really Matter Whether Students' Contributions Are Spoken versus Typed in an Intelligent Tutoring System with Natural Language?

Sidney K. D'Mello, Nia Dowell, and Arthur Graesser

University of Memphis

Author Note

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

We gratefully acknowledge the associate editor Dr. Daniel Morrow and two reviewers whose insightful and useful suggestions substantially improved this paper.

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

Contact person: Sidney D'Mello, 202 Psychology Building, University of Memphis, Memphis, TN 38152. Phone: 901 378-0531. Fax: 901 678-1336. Email: sdmello@memphis.edu

Abstract

There is the question of whether learning differs when students speak versus type their responses when interacting with intelligent tutoring systems with natural language dialogues. Theoretical bases exist for three contrasting hypotheses. The *speech facilitation* hypothesis predicts that spoken input will *increase* learning, whereas the *text facilitation* hypothesis predicts typed input will be superior. The *modality equivalence hypothesis* claims that learning gains will be equivalent. Previous experiments that tested these hypotheses were confounded by automated speech recognition systems with substantial error rates that were detected by learners. We addressed this concern in two experiments via a Wizard of Oz procedure, where a human intercepted the learner's speech and transcribed the utterances before submitting them to the tutor. The overall pattern of the results supported the following conclusions: (a) learning gains associated with spoken and typed input were on par and quantitatively higher than a no-intervention control, (b) participants' evaluations of the session were not influenced by modality, and (c) there were no modality effects associated with differences in prior knowledge and typing proficiency. Although the results generally support the modality equivalence hypothesis, highly motivated learners reported lower cognitive load and demonstrated increased learning when typing compared to speaking. We discuss the implications of our findings for intelligent tutoring systems that can support typed and spoken input.

Keywords: speech facilitation, text facilitation, modality equivalence, intelligent tutoring systems, spoken dialogues

Does It Really Matter Whether Students' Contributions Are Spoken versus Typed in an Intelligent Tutoring System with Natural Language?

Over the last few decades, Intelligent Tutoring Systems (ITSs) have emerged as valuable tools for promoting learning by either implementing ideal tutoring strategies or by attempting to simulate human tutoring (Psotka, Massey, & Mutter, 1988; Sleeman & Brown, 1982; Woolf, 2009). Some of the ideal tutoring strategies that ITSs implement are 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, Douglass, & Qin, 2005; Gertner & VanLehn, 2000; Koedinger, Anderson, Hadley, & Mark, 1997; Lesgold, Lajoie, Bunzo, & Eggan, 1992; Sleeman & Brown, 1982). 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, 2001; 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 (Cohen, Kulik, & Kulik, 1982).

One subset of these ITSs implements natural language dialogue that is comparable to the conversations that occur in human tutoring. These include AutoTutor (Graesser, et al., 2004; VanLehn, et al., 2007), why-Atlas (Graesser, VanLehn, Rose, Jordan, & Harter, 2001; VanLehn, et al., 2002), CIRCSIM-Tutor (Shah, Evens, Michael, & Rovick, 2002), DC-Trains (Pon-Barry, Clark, Schultz, Bratt, & Peters, 2004), and Mission Rehearsal (Gratch & Marsella, 2001). These different computer tutors vary in the extent to which they simulate human dialogue mechanisms,

but all of them attempt to comprehend natural language, formulate adaptive responses, and implement pedagogical strategies to help students learn.

There is a question of how the input modality of the student affects the process of active knowledge construction. This is an important question because any ITS with natural language dialogue needs to make a commitment on the input modality of student contributions. Should students type in the information in their conversational turns? Or, should they produce spoken contributions? Most ITSs with natural language dialogue have supported only typed input, but an increasing number of systems have been implementing spoken tutorial dialogues (D'Mello, King, & Graesser, in press; Litman, et al., 2006; Mostow & Aist, 2001; Pon-Barry, et al., 2004). The hope is that speech enabled ITSs will more closely mirror human-to-human communication and thereby significantly increase learning gains and engagement. As advances in automatic speech recognition (ASR) technologies continue to accrue, the next generation of ITS developers will have to face the difficult decision of whether to implement spoken tutorial dialogues as an alternative to conventional typed dialogues. This is a non-trivial design decision with important consequences, so there needs to be empirical evidence to provide guidance.

Theoretical Predictions for Input modality Effects

Theoretical bases exist for three contrasting predictions regarding the effectiveness of spoken dialogues for enhancing learning: (1) spoken dialogues will *decrease* the efficacy of ITSs (text facilitation hypothesis), (2) spoken dialogues will *increase* the efficacy of ITSs (speech facilitation hypothesis), and (3) spoken dialogues will *have no effect on* the efficacy of ITSs (modality equivalence hypothesis). These hypotheses can be understood from the broad perspective of constructivism, which is a theoretical framework adopted by many researchers who are exploring deeper levels of comprehension (Biggs, 1995; Bransford, Goldman, & Vye,

1991; Brown, 1988; Chi, Deleeuw, Chiu, & Lavancher, 1994; Palinscar & Brown, 1984; Piaget, 1952; Rogoff, 1990; VanLehn, Jones, & Chi, 1992; Vygotsky, 1978). Constructivist approaches have shaped the standards for curriculum and instruction in the United States during the last decade. According to this approach, the learner needs to actively construct coherent, explanation-based meanings and knowledge by interacting with the world and other people. Learning environments should stimulate active construction of knowledge and provide feedback and explanations on these constructions rather than being mere information delivery systems.

ITSs that adhere to constructivist principles attempt to get learners to do most of the talking by pumping for information, providing hints, prompts, forced choices, and other pedagogical scaffolds. The onus of knowledge construction is placed on the learner and involves cognitive processes such as perception, management of working memory, planning, the production of language and discourse constituents, and the consolidation of subject matter knowledge. The differential merits of spoken versus typed input can be understood by considering how *input modality* affects these cognitive processes that are presumably linked to deeper comprehension and learning gains.

In general, there is a tradeoff between the ease of a student producing a response in a conversational turn (called a student contribution) and the quality of the student contribution. Spoken language is of course easier to produce whereas typed input affords higher quality. The student's typed contribution remains on the screen for them to refer to whereas spoken utterances are evanescent, i.e., they disappear shortly after they are spoken (Clark, 1996; Clark & Brennan, 1991; Gergle, Millen, Kraut, & Fussell, 2004). According to the text facilitation hypothesis, the persistence of the student's text on the screen affords added perceptual processing, rereading,

and memory encoding, thereby increasing learning gains if students use these textual representations to process the material more deeply.

When students type in their input and see what they type, they can evaluate their responses, remove errors or misspellings, and revise their contributions. The additional time to reflect on their composition is expected to yield superior learning gains compared to speech-based ITSs in which students have to generate their responses in real time (Quinlan, 2004). Of course, added time to provide a response is expected to promote deeper learning to the extent that students use this time to reflect and improve their contributions.

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, so learning should benefit from the planning that occurs in typed contributions.

Cognitive load theory (Chandler & Sweller, 1991; Sweller, 1988) also suggests that typed dialogues will be superior to spoken dialogues. According to cognitive load theory, individuals have a limited working memory so they are able to focus attention and processing on a limited amount of information during any given time span. During the composition of lengthy typed contributions, cognitive load can be effectively managed by outsourcing information into an external record for reinspection, thereby allowing more resources to be allocated to thought and problem solving (but see below for non-proficient typists).

In summary, the *text facilitation hypothesis* posits that typed input will enjoy advantages over the speech input because typed input affords the opportunity to reinspect the text, enhance memory encoding, make revisions, engage in enhanced planning and composition, and reduce cognitive load.

The second hypothesis, the *speech facilitation hypothesis*, makes a very different prediction, namely that spoken dialogues will facilitate learning over typed input. One reason is the comparative expression gap between thought and language. The gap is smaller between thought and speech than thought and written text (Chafe, 1982; Tannen, 1982). Hence, spoken responses are relatively effortless and natural when compared to written responses. Because of the ease of spoken contributions, the volume of content is typically longer when responses are spoken than typed. Given that learning is correlated with the volume of contributions by the students (Litman, et al., 2006), following the constructivist framework, it would be predicted that spoken contributions 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 ITS we tested (AutoTutor, which will be 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 need to constantly shift their attention from the image to the text-box when constructing a typed response; this can potentially increase cognitive load. This problem can be mitigated to some extent by students speaking their contributions instead of typing them. The student can focus on aspects of the image while simultaneously constructing their spoken responses, thereby reducing attentional demands.

There is another 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. Mayer's claims apply to students perceiving and comprehending input, but the advantages of the spoken dialogues might similarly hold for students' production of information while interacting with an agent that speaks. Spoken instruction may prime the social rules of human-to-human communication, resulting in a spoken interaction mindset that encourages active processing of incoming material by learners. Ideally the spoken conversation between student and tutor would have short latencies between turns. Such advantages may be compromised in communication media that involve pauses between turns. The conversations in the present study do involve a delay of a few seconds in conversational turns between the student utterances and the tutor responses, so this advantage may be modest or nonexistent.

In summary, there are distinct reasons to expect an input modality effect when learners speak or type their responses to the tutor. Table 1 summarizes the hypothesized benefits of each input modality with respect the cognitive processes they influence and their impact on learning gains. As evident from Table 1, the extent to which one input modality will promote learning compared to the other should depend upon the cumulative benefits of one modality over the other.

It is important to note that the advantages of typed and spoken input listed in Table 1 depend upon the degree to which the learner is actively engaging in processing the material. For example, the availability of a written record on the screen is an advantage of typed input because it provides an opportunity for additional encoding. However, shallow rehearsal of the information is not expected to impact learning (Craik & Lockhart, 1972; Craik & Tulving, 1972). Learning should only be affected if learners engage in deeper processing activities by using the external record to revise, plan, reinspect, and improve their responses.

INSERT TABLE 1 ABOUT HERE

According to the *modality equivalence hypothesis*, input modality will have little to no impact on learning gains when the two modalities are approximately equivalent in terms of communication efficiency and expressivity. One reason to expect a null effect for input modality is that the benefits of each modality cancel each other out. Specifically, Table 1 lists four potential benefits for speech input and four for typed input. Assuming that each factor has equivalent effects on learning, then the combined effect will result in equivalent learning.

There is also a cost-benefit trade-off *within* a single modality. For example, we have suggested that typing can potentially alleviate cognitive load because the learner can outsource information to the screen instead of maintaining an active mental representation of a lengthy response. However, typing can also increase cognitive load when images are present because the learner has to shift attention from the input box to the image while constructing a response. Alternatively, the hypothesized reduction of cognitive load when typing (due to availability of an external record) can potentially be nullified when non-proficient typists have to struggle to produce a typed response, thereby increasing cognitive load. Hence, one explanation for the modality equivalence hypothesis is that the associated costs and benefits cancel each other out both *across* and *within* modalities.

An alternate explanation for modality equivalence is that the *content* is more important than the *medium* of communication (Graesser, et al., 2003). Simply put, the medium is not the message; rather, the message is the message. This hypothesis would predict equivalent learning gains for both conditions to the extent that the content is the same. In essence, it is the dialogue between the human and the tutor, as well as the content covered, that explains the learning gains (Whittaker, 2003). As the content of tutorial dialogue emerges, there is a rich conversation

history that covers meaning, world knowledge, pragmatic aspects of communication, and social dynamics between the tutor and learner. This rich content eclipses any differences there might be between the typed and spoken modalities.

Previous Research Comparing Spoken and Typed Contributions

Available experiments that have directly compared spoken versus typed tutorial dialogues have interestingly produced different results. Litman et al. (2006) reported two experiments examining differences between typed and spoken dialogues. The first experiment compared typed versus spoken dialogues with a human tutor whereas the second experiment compared modalities with the ITSPOKE computer tutor (Litman & Silliman, 2004). Their results indicated that changing the input modality from typed to spoken dialogues had a substantial positive impact on learning in the experiment with the human tutor but had no significant impact in the experiment with the computer tutor. Although there were advantages to spoken tutorial interventions, the benefits apparently applied only to human-human communication settings, not to human-computer interactions.

Nevertheless, there is one potential concern with the Litman human-computer experiment that might explain the failure to replicate the findings of the human-human experiment. The text of the student's spoken turns was not displayed to the student in the human-human experiment. In contrast, the tutor in the human-computer study presented a dialogue history of the student's spoken responses, so the students were aware of speech recognition 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.

This concern was addressed in a recent study with a different ITS, called AutoTutor (D'Mello, et al., in press). This study did not give students access to a dialogue history, so they were not explicitly aware of speech recognition errors. Nevertheless, there was no statistically significant difference in learning gains between the typed and spoken conditions[1].

Although errors were not explicitly displayed to the students in the D'Mello et al. (in press) study, it is still possible that they could have inferred system failures due to erroneous speech recognition. In fact, participants' responses to items on a post-interaction questionnaire indicated that AutoTutor understood significantly fewer of their spoken responses than their typed responses. Such student biases towards spoken human-computer interactions might negatively affect their ability to learn. The speech recognition system utilized in ITSPOKE had a word error rate[2] of 31.2%. The AutoTutor system had a word error rate of 53.2%. Since students who interacted with both systems were aware of these errors, this in turn may have lowered student expectations, confidence, and motivation towards the tutoring system. A negative impact on learning may have resulted from a lack of confidence in an ITS that had such a mismatch of expectations. The superior learning gains obtained from spoken dialogues with human tutors might be replicated in ITSs if speech recognition accuracy was equilibrated.

Another important aspect that was neglected in the previous research pertains to the role of individual differences in moderating the (hypothesized) effects of input modality on learning.

---

[1] There was a .23 sigma effect in favor of the typed condition. This effect might have been significant with a larger sample.

[2] Word error rate (*WER*) is a standard metric for assessing the reliability of automatic speech recognition systems. It is computed from the number of substitutions (*S*), deletions (*D*), and insertions (*I*) in the automatically recognized text (with errors) when compared to the ideal text (no errors) of *N* words.

For example, people are presumably not used to speaking aloud to computers and some individuals might be uncomfortable with this activity. This potential dislike of spoken interfaces might mitigate some of the hypothesized benefits of spoken responses and these individuals might learn more when they type their responses. Alternatively, as mentioned above, non-proficient typists might prefer to speak their responses and might even learn less when typing due to the increased cognitive load associated with producing typed responses.

Individual differences in motivation, interest, and knowledge are also expected to have an impact on how input modality influences learning (Berlyne, 1978; Pekrun, Goetz, Daniels, Stupnisky, & Raymond, 2010; Tobias, 1994). Input modality is only expected to facilitate learning if learners are actively processing the material and are producing effortful responses. An unmotivated learner with minimal interest in the task is unlikely to invest substantial cognitive effort in learning, thereby nullifying any of the hypothesized modality effects. It is also possible that prior knowledge might moderate input modality effects because the more knowledgeable learners might be oblivious to modality differences; these learners know the answers and simply deliver them to the tutor while speaking or typing.

In summary, there are two concerns pertaining to the previous research that tested input modality effects on learning. First, there is the concern that the speech recognition errors in the previous experiments by Litman and D'Mello might be a confounding variable in testing the effects of spoken and typed input. Second, the fact that the previous experiments did not assess the impact of individual difference effects is problematic because typing proficiency, prior experience with spoken interfaces, prior knowledge, and motivation to learn are individual differences that might moderate the effects of input modality on learning.

These concerns were alleviated in two experiments that tested the speech facilitation, text facilitation and modality equivalence hypotheses while students learned computer literacy with AutoTutor. While the earlier experiments used a commercially available automatic speech recognition system for speech to text transcription, the current experiment used a human to intercept participants' spoken utterances and manually transcribe the speech before submitting it to the AutoTutor system. This method yielded accurate speech recognition and participants were unaware that speech recognition in this experiment was being performed by a human. In this fashion, we were able to assess the effects of the modality on learning gains without the confound of imperfect speech recognition. Relevant individual difference measures were also collected in Experiment 2, thereby providing us with the ability to test whether these differences moderated the effect of input modality on learning.

Experiment 1

Method

Design. The experiment had a repeated-measures design which consisted of a pretest, a tutorial session with one modality, a questionnaire, a tutorial session with the other modality, a questionnaire, and a posttest. Therefore, each student participated in two computer literacy tutorial interactions, one with speech-based input and one with typed input. The order in which students used these input methods (speech first, typed second or typed first, speech second) was counterbalanced across participants. All participants were also assigned to a control condition; for this condition they received no tutorial instruction.

Assignment of the computer literacy topics (hardware, operating systems, Internet) to tutor conditions was counterbalanced across all participants using a 3 by 3 Latin square. For example, a given participant might have been assigned the hardware topic in the speech

condition, operating systems in the typed condition, and Internet in the control condition, while another participant might have been assigned hardware in the typed condition, operating systems in the control condition, and Internet in the speech condition.

It is important to clarify a couple of important points regarding the no-intervention control condition. From the example above, learning gains for the control condition would be computed as the difference between posttest and pretest scores (described below) for questions pertaining to the Internet topic because the learner did not receive any tutoring for this topic. The purpose of the control condition was to provide a comparison condition to assess whether any significant learning occurred in the spoken and typed conditions compared to the no-intervention control. This design might be confounded with carry-over effects from the spoken and typed sessions onto the control condition, but previous studies have reported that no such carry over effects occur in these repeated measures designs that test AutoTutor on particular topics and that such designs yield the same results as between-subjects designs (Graesser, et al., 2004; VanLehn, et al., 2007).

Participants. Twenty-four undergraduate students from a mid-south university in the United States participated in the experiment for course credit. Data from one participant was eliminated from the experiment due to experimenter error. There were 15 participants who scored at or below the median on the pre-test (described below), while the remaining 8 participants scored above the median. The median pretest score was 0.25, which is what can be expected by chance given that the test consisted of four-alternative multiple-choice questions.

AutoTutor. The tutoring system used in the experiment was AutoTutor, a fully automated tutor for Newtonian physics, computer literacy, and critical thinking (Graesser, Jeon, & Dufty, 2008; Graesser, et al., 2004). AutoTutor presents students with a series of challenging

problems (or main questions), each requiring approximately three to seven sentences of information for a correct answer. Examples of main questions are, "How can John's computer have a virus but still boot to the point where the operating system starts?" (hardware question), "How would you design an operating system that can manage memory demands from multiple concurrent jobs?" (operating systems), and "How will you design a network that will continue to function, even if some connections are destroyed?" (Internet).

In order to correctly answer a main question, students need to articulate 3-7 sentential expressions, which normally involves between 25 and 100 turns in a conversation. When presented with these questions, students typically respond with answers that are only one word to two sentences in length. In order to guide students in their construction of an improved answer, AutoTutor actively monitors students' knowledge states and engages learners in a turn-based dialogue. AutoTutor adaptively manages the tutorial dialogue by providing feedback, pumping the learner for more information, giving hints, correcting misconceptions, answering questions, and summarizing answers. Using these strategies, AutoTutor adheres to constructivist theories of pedagogy (Biggs, 1995; Chi, Roy, & Hausmann, 2008; Moshman, 1982) by allowing students to chart their own course through the tutorial dialogue and to build their own answers to difficult questions.

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, Lu et al., 2004; Graesser et al., 2005, 2008). Expectations and misconceptions form the underlying conceptual structure that drives AutoTutor's dialogue and are the primary pedagogical methods of scaffolding good student answers. Both AutoTutor (Graesser et al., 2005) 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 four expectations below 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,

(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 & Person, 1994; VanLehn et al., 2007). The 5 steps of the dialogue frame are: (1) Tutor asks main question, (2) Student gives initial answer, (3) Tutor gives short feedback on the quality of the student's answer in #2, (4) Tutor and student collaboratively interact via expectation and misconception tailored dialogue, and (5) 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 either positive (e.g. "very good", "bravo"), negative (e.g. "not quite", "almost"), or neutral (e.g. "uh huh", "okay"). The second slot advances the coverage of the ideal answer with either 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 for the floor to shift from AutoTutor as the speaker 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.

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 drive the comparison whereas pattern completions drive the selection of AutoTutor hints and prompts. These matching and completion operations are based on symbolic interpretation algorithms (Rus & Graesser, 2007) and statistical semantic matching algorithms (Landauer et al., 2007). These include an Inverse Weighted Word Frequency Overlap algorithm, Latent Semantic Analysis, and a Speech Act Classifier. Details on the actual mechanisms that AutoTutor uses to interpret the learner's contributions are presented in previous publications (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007).

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 adaptively selects those hints and prompts that fill missing constituents and thereby achieves pattern completion. Empirical validation of the conversational smoothness of AutoTutor's dialogue has been presented in previous publications (Graesser, et al., 2003; Person & Graesser, 2002).

Table 2 illustrates an excerpt conversation with AutoTutor that was extracted from an actual tutoring session. This session was with a relatively verbose, knowledgeable student about computer literacy so the conversation is comparatively short.

INSERT TABLE 2 ABOUT HERE

AutoTutor's interface has five major windows shown in *Figure 1*. Window 1 (top of screen) is the main question that stays on the computer screen throughout the conversation for the question. Window 2 (left middle) is the animated conversational agent that speaks the content of AutoTutor's turns. Window 3 (right middle) is either blank or has auxiliary diagrams. Window 4 (right bottom) displays the students' answers as they type it in. Window 5 (left bottom) displays the dialogue history of the student and the tutor. However, Window 5 was deactivated in this experiment so students had no access to the dialogue history (discussed below).

INSERT FIGURE 1 ABOUT HERE

Two versions of AutoTutor were used in the current experiment. These included AutoTutor with typed input and AutoTutor with spoken input. The typed-input version was the traditional AutoTutor system. Participants typed their responses into a text area (bottom right in *Figure 1*) and pressed the *Enter* key to submit their response to the tutor. This allowed students to view and edit their typed responses before submitting them to AutoTutor. The submission pane displayed only the text that participants were currently constructing. Once the text was submitted to the tutor, it was no longer available for participants to view on the screen.

In the speech-input version students spoke their responses through a microphone and were unable to see the text of their responses. Participants' speech and the audio generated by the AutoTutor animated conversational agent were recorded for offline analyses. In this version

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.

Instead of an automatic speech recognizer, a human in an adjacent room listened to participants' speech and manually transcribed the utterances. The human started transcribing when the participant pressed the *F1* key and stopped transcribing when the participant pressed the *F2* key. Participants were not aware that their speech was being recognized by a human instead of the computer.

It is important to highlight three 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 1*) was not available for the participants to view in both conditions. The rationale behind this decision was to prevent participants from viewing any errors made by the human transcriber.

Second, in both versions, an embodied pedagogical agent (see middle left of *Figure 1*) delivered all of the tutor's dialogue via synthesized speech. So 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.

Third, the human transcriber started typing immediately when the participant pressed the *F1* key to initiate a response and stopped typing three seconds after participants pressed the *F2* key, which signified the end of the response. Preliminary testing indicated that a three-second lag was sufficient to transcribe the entire spoken response because student contributions in one turn typically consist of only a few words. AutoTutor's response in the *typed* version was also delayed by three seconds via a software timer. Hence, AutoTutor's response latency (time between answer submission and tutor's response) was equivalent in both versions.

Knowledge Tests. Participants were tested on their knowledge of computer literacy topics both before and after the tutorial session (pretest and posttest, respectively). Based on these test scores, learning gains were computed to determine the amount of knowledge that students acquired during the tutorial session. The testing materials were adapted from computer literacy tests used in previous experiments involving AutoTutor (Graesser, et al., 2004), and were comprised of questions that assessed students' knowledge of all three computer literacy topics. Each test contained 24 multiple-choice questions: 8 questions on hardware, 8 questions on operating systems, and 8 questions on the Internet. Participants completed alternate test versions for pretest and posttest. The two test versions, composed of different questions, tested learners on the same subject matter and content. The assignment of test versions to pretest versus posttest was counterbalanced across participants.

The 4-alternative multiple-choice format was designed to assess deep levels of knowledge. The questions required answers that involved inferences and deep reasoning, 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, in press). Example questions from the knowledge tests are presented in the Appendix A.

Post-interaction Questionnaire. The post-interaction questionnaire asked participants to evaluate their tutorial session on measures of perceived performance, user satisfaction, and task difficulty. Each participant completed the questionnaire at two points during the experiment. It was administered once after the speech-based interaction and once after the text-based interaction. The questionnaire consisted of 8 items.

The first 6 out of 8 questions required participants to rate the following 6 statements on a 6-point scale: 1) "I enjoyed interacting with AutoTutor," 2) "I felt that my interaction with AutoTutor was comparable to an interaction with a human tutor," 3) "I felt that AutoTutor did *not* understand what I said," 4) "I felt engaged during the tutoring session," 5) "I felt that AutoTutor was difficult to use and work with," and 6) "I felt that I learned new information from AutoTutor." The participants were instructed to respond by choosing one of 6 alternatives: *strongly disagree*, *disagree*, *somewhat disagree*, *somewhat agree*, *agree*, and *strongly agree*. Based on their response, participants were assigned a score of 1 (*strongly disagree*) to 6 (*strongly agree*).

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: "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: "Apart from the content of the tutoring session, how difficult was it to learn new information from the tutor?" Participants were provided with 7 response alternatives: *very easy*, *somewhat easy*, *slightly easy*, *medium*, *slightly hard*, *somewhat hard*, and *very hard*. Based on their response, participants were assigned a score of 1 (*very easy*) to 7 (*very hard*).

Procedure. Participants were tested individually during a two-hour session. First, participants completed an informed consent and then the pretest. Participants were instructed to

take a seat at the computer console and to put on a pair of headphones that included a microphone. Next, the general features of AutoTutor's dialogue and pedagogical strategies were described to the learner. On the basis of random assignment, learners were either tutored through speech-based input during their first interaction (speech-first group) or through typed input during their first interaction (typed-first group). Before the first tutorial interaction, participants were led through the voice training period on an automated speech recognition system. This provided participants with the illusion that their speech was being recognized by a computer, when in fact, it was being recognized by a human. Participants then interacted with one version of AutoTutor until 4 main questions were successfully answered or the training period had elapsed. The training period was 40 minutes long. After the tutorial interaction, participants were asked to complete the post-interaction questionnaire.

Participants were subsequently informed that they would be using a different input method during the second tutorial interaction. The procedure for the second tutorial interaction was the same as in the first tutorial interaction, with the exception that the interaction modality was changed. Finally, participants in both groups completed the posttest and were debriefed.

Speech Recognition Accuracy. The average length of a spoken utterance was 24.4 characters (*SD* = 13.0), which was not statistically different from the average length of a typed utterance (*M* = 28.0, *SD* = 15.8). Hence, input modality had no significant impact on response verbosity.

Speech recognition performance of the human transcriber was measured using word error rate (*WER*). In order to measure *WER*, the automatically recognized text of each student dialogue turn was compared to a manual transcription of that dialogue turn. The manual transcription was prepared by an experimenter from the audio recordings of the tutorial interactions. The human

experimenter transcribed a random sample of 126 utterances from the corpus of 1230 spoken utterances. An equal distribution (approximately 10%) of spoken utterances from all 23 participants was included in the random sample.

WER was computed separately for each utterance and averaged across utterances yielding an aggregate score for each participant. As expected, speech recognition was almost perfect and the error rate was substantially lower (M = .110, SD = .175) than what was achieved in the earlier experiment reported by (D'Mello, et al., in press) with automatic speech recognition (M = .532, SD = .346), d = 1.54.

Results

Our analyses were organized around two research questions. First, how did learning gains compare in the no-intervention control, the spoken condition, and typed condition? Second, did participants prefer speaking or typing their responses? Participants' scores on the knowledge tests were used to answer the first question, whereas their responses on the post interaction questionnaires were analyzed to answer the second question.

Learning Gains. The pretests and posttests were scored for the proportion of questions (out of 24) that participants answered correctly. The measure of learning consisted of proportional learning gains, computed as: (posttest scores – pretest scores) / (1 - pretest scores).

A 2 × 3 repeated-measures ANOVA was performed on proportional learning gains, with *condition* (speech, typed, control) as a within-subject factor and *interaction-order* (speech-typed versus typed-speech) as a between subject factor. The main effect for interaction-order was not significant, nor was the condition × interaction-order effect. Therefore, speaking first and then typing versus typing first and then speaking did not impact learning gains.

The proportional learning gains showed a significant difference among conditions, $F(2, 42) = 3.90$, $p < .05$, $MSe = .114$, partial $\eta^2 = .157$. Planned comparisons were performed to test whether: (1) significant learning occurred in the spoken and typed conditions compared to the no tutor condition (a one-tailed test), and (2) whether there was a significant difference between speaking and typing (two-tailed test). As expected on the basis of theory and previous studies (Graesser, et al., 2004), a one-sample *t*-test confirmed that learning gains for the no tutor condition was equivalent to 0 ($p = .307$).

Learning gains for both the spoken ($M = .320$, $SD = .329$) and typed ($M = .268$, $SD = .372$) conditions were higher than the no tutor condition ($M = .059$, $SD = .271$), $t(22) = 2.66$, $p < .05$, $d = .85$ for speech and $t(22) = 2.06$, $p < .05$, $d = .66$ for typed. There was no statistical difference between the spoken and typed conditions, $p = .675$, $d = .12$.

ITSs such as AutoTutor are generally more effective for low-domain knowledge learners compared to participant who have some prior knowledge on the tutorial topic (VanLehn et al., 2007). It might be the case that the effect of input modality might interact with prior knowledge. However, our sample mainly consisted of low-domain knowledge students, so we performed a follow-up analysis on the 15 learners who scored below the median on the pretest. Learning gains for the spoken and typed conditions were statistically equivalent, $M = .374$ ($SD = .336$) for speech and $M = .264$ ($SD = .324$) for typed, $p = .308$, $d = .27$.

There is the concern that the lack of significant differences between the spoken and typed conditions might be attributed to our small sample size. The effect size between the spoken and typed condition was .12 sigma, which is consistent with a very small effect (Cohen, 1992). A power analysis indicated that a sample size of 536 participants would be required to detect this effect (power = .8 and $\alpha = .05$) with a two-tailed test. Aside from this qualification, in our view,

it is difficult to attribute the lack of statistical significance between the spoken and typed conditions to merely sample size. Our sample size was adequate to detect the .8 sigma mean effect size obtained from previous studies with AutoTutor as well as the 1.0 sigma effect size obtained by other ITSs (Corbett, 2001; Corbett, et al., 1999; Graesser, et al., 2004; VanLehn, et al., 2007). Our sample was also sufficiently large to detect the .74 sigma effect reported by Litman et al. (2006). Although relying on the .74 effect might be problematic due to the potential confounds in the Litman study, it should be noted that our sample was sufficiently large to detect a medium effect of .54 sigma or higher.

Another concern pertains to our use of a within-subjects experimental design in lieu of a between-subjects design. We adopted this design because individual difference variability is partialled out in a within-subjects design. This not only makes it more likely to detect a subtle effect, but also allows us to track individual differences. For example, perhaps some participants benefit primarily from spoken contributions and others typed contributions. The obvious concern with a within-subjects design is that participating in the second session on a related subject matter might be influenced somehow by the first session. We performed a follow up analysis to address this concern by considering only the first AutoTutor session for each participant. Comparisons between the typed and spoken conditions yielded a non-significant small effect of .29 sigma, with 11 to 12 participants in each group.

Subjective Evaluation of Tutorial Session. We compared participants' ratings on the post-interaction questionnaires that were administered after each tutorial session (see Table 3). We performed separate ANOVAs on each of the 8 questions to determine whether participants preferred the spoken or typed versions of AutoTutor. There was a 2 × 2 (input modality: spoken versus typed × interaction order: spoken-typed vs. typed-spoken) ANOVA conducted for each

question. Input modality was a within-subjects factor whereas interaction order was a between-subjects factor.

The results indicated that the input modality main effect was not significant for any of the questions. Thus, the input modality had no impact on learners' perceptions of the tutoring session. The main effect for interaction order was never significant nor were the interactions between interaction order and input modality. Therefore, participants' subjective evaluations of the tutorial session were not influenced by modality and by typing first versus speaking first.

INSERT TABLE 3 ABOUT HERE

Discussion

It is tempting to conclude that the results from Experiment 1 favor the modality equivalence hypothesis because we failed to find an input modality effect. However, before we accept this conclusion too cavalierly, it is important to consider whether any methodological factors prevented us from finding an effect. It appears that there are three potential concerns with Experiment 1 that might have contributed to the null effect.

The first concern pertains to the length and nature of student responses to tutor questions. Although we hypothesized that speaking would yield longer responses than typing, this was not confirmed in the experiment. Whether students spoke or typed, their responses were very short, generally ranged from two to six words (mean = 4.4). The responses usually consist of single word answers to prompts or word fragments with syntactic structures that were either very simple or ungrammatical (e.g., "It is storage", "because it is fast"). Student responses to pumps (e.g., What else?) suffered the same problems, even though tutor pumps were explicitly designed to extract as much information as possible from the student. These problems with student responses have minimal impact on AutoTutor because its natural language understanding

mechanisms do not require verbose, syntactically well-formed contributions. However, many of the presumed benefits of the input modalities on constructs such as working memory load, enhanced perception, efficiency of speech production, planning, and consolidation become less significant when the responses consist of a few words. For example, we hypothesized that speech is beneficial because if affords the efficient production of utterances, while typing afford students the ability to inspect and revise their responses. It is unlikely that these modalities have a noticeable impact when the average contribution is only four words.

Another related concern that might have mitigated the discovery of a modality effect is the fragmented nature of AutoTutor's dialogue. That is, the design of AutoTutor's discourse management system, which is based on human tutoring, may have encouraged shorter student responses that lacked coherent threads of reasoning. The problems stem from the fact that the ordering of the four expectations (listed in the AutoTutor section of the Methods) are not covered in a fixed order (i.e., E1, E2, E3, E4) during the dialogue. Instead, the ordering of the expectations is dynamically decided for each learner on the basis of their zone of proximal development (Brown, Ellery, & Campione, 1998; Vygotsky, 1978).

Some elaboration of the dialogue management mechanism should clarify why student contributions were short and had moderate cohesion. Each problem (or main question) involves the following major steps in the dialogue planner.

1. Present main question and get a response from the student
2. Pump student for more information (e.g., "What else?").
3. Compute the semantic match between the aggregated student responses from #1 and #2 and each of the expectations for the main question. The expectation with the highest semantic match is posted as the *current expectation*.

4. Provide hints, prompts and assertions in order to get the student to articulate the content of the current expectation. This step occurs over multiple turns until the expectation is covered.
5. Aggregate the student's responses from #1, #2, and #3 and compute the semantic match between this aggregated response and the remaining expectations (i.e., expectation not yet covered). Select the uncovered expectation with the highest semantic match to the student's aggregated response. Repeat #4 and #5 until all the expectations are covered.
6. Give a summary that provides the content of all the expectations for the main student.

This expectation selection mechanism can be considered to be a form of knowledge tracing (Corbett & Anderson, 1994) because it dynamically selects a sub problem (i.e., expectation to cover) on the fringe of each student's zone of proximal development. However, one presumably negative side effect of this mechanism is that it yields fragmented dialogues with no cohesive thread connecting the different expectations. These fragmented dialogues run the risk of confusing and possibly frustrating students, especially when they are unsure about how to proceed. This might be problematic for the current study because if students get so lost while in the throes of AutoTutor dialogue, then input modality loses its impact.

In summary, it is possible that we did not find an effect in Experiment 1 because AutoTutor did not provide students with the opportunity to capitalize on the merits of the different modalities. Additionally, as discussed in the Introduction, it might also be the case that input modality has differential impacts for different students. Experiment 1 did not include any individual difference measures, thereby precluding us from testing these hypotheses. An effect

for input modality might be more apparent for some students when the dialogue is more cohesive and when students are required to generate substantive content by composing essays, summarizing topics, and providing lengthy responses. These concerns were addressed in a follow-up study that utilized a modified version of AutoTutor and collected relevant individual difference measures.

Experiment 2

Method

Thirty-six undergraduate students from a mid-south university in the United States participated in the experiment for course credit. One participant was excluded from the study due to a computer malfunction. There were 16 participants who scored below the median (0.25) on the pre-test, while the remaining 19 participants scored above the median.

The design, knowledge-tests, and procedure for Experiment 2 were identical to Experiment 1, with three exceptions. First, participants completed a one-minute typing test at the start of the experiment. The words typed per minute (WPM) was recorded upon completion of the test. We included this test in order to investigate whether typing ability was associated with an input modality effect.

Second, participants completed a Computer Usage and Knowledge Questionnaire. This questionnaire consisted of 11 items that measured participants experience with computers, impressions of computers, confidence that they can learn computer literacy, prior exposure to spoken interfaces, and other related measures (see Appendix B). Some of the items on this questionnaire (1, 3, 4, 5, 6, 7, 8) were adapted from an existing questionnaire ("PC Usage Questionnaire," 2010), while the remaining items were created for the purpose of this study.

The third difference between experiments pertained to the version of AutoTutor. The version used in Experiment 2 was explicitly designed to address concerns about the version of AutoTutor used in Experiment 1 (discussed above). The logs of the new version of AutoTutor were augmented to record a number of interaction parameters, such as planning time, response time, and several measures of verbosity. While participants in Experiment 1 submitted their typed responses by pressing the *Enter* key, participants in Experiment 2 were required to press the *F1* and *F2* keys to start and stop typing (as well as speaking) in order to accurately compare planning time across modalities. In order to alleviate problems related to fragmented dialogues with modest cohesion, AutoTutor's dialogue planning mechanism was replaced with a simpler mechanism that iterated through each expectation in a predefined sequence. The experimenters verified that the generated dialogues were internally cohesive (i.e., cohesion across turns) prior to data collection.

The new version of AutoTutor adopted a number of measures to increase the verbosity of students' responses. In particular, students were not permitted to advance past the main question and the pump unless they provided a minimum of six words. Failure to provide the requisite number of words led to a gentle prompt to get them to articulate more content (e.g., "Come on. I need you to say more than that in order for me to tutor you. Please try again"). The system advanced the dialogue if students still neglected to provide sufficient content because preventing them from advancing induced substantial frustration.

While the original AutoTutor provided a summary at the end of each problem, the new version prompted students to provide a summary of at least 15 words before delivering its own summary. There was also a difference in how hints and prompts were used. The original tutor used both hints and prompts to get the student to articulate an expectation whereas the new tutor

never prompted the student. One or two-word answers are sufficient to answer a prompt, which was incompatible with our goal to increase student verbosity in Experiment 2.

There is the concern that the increased verbosity of the student responses would create some difficulties for the human transcriber who had to respond in real-time. However, the quality of the transcription had no impact on AutoTutor's dialogue plan because, as described above, the new version was not dynamically adaptive to individual students with respect to selection of expectations, hints and prompts. However, transcription errors would impact the short feedback (positive, neutral, negative) that AutoTutor expresses after student turns. For example, AutoTutor might end up giving more negative feedback then it should if the transcriber misses some of the content words in the spoken response. In order to alleviate this problem, AutoTutor's feedback mechanism was modified to provide neutral feedback (e.g., "ok", "alright") when students provided a verbose response (i.e., initial answers main questions, answers to AutoTutor pumps, and student summaries) irrespective of whether they spoke or typed their contributions. Accurate formative feedback was provided by AutoTutor after students responded to hints because these responses were the approximately the length of a short sentence or clause and transcribers could enter these accurately.

Results

Learning Gains. The learning gains were analyzed in the same manner as in Experiment 1. A 2 × 3 (*condition* × *interaction-order*) repeated-measures ANOVA yielded a significant main effect for condition, $F(2, 58) = 11.53$, $MSe = .095$, $p < .001$, partial $\eta^2 = .284$. Learning gains for the spoken ($M = .277$, $SD = .349$) and typed ($M = .343$, $SD = .350$) condition were significantly higher than the no tutor condition ($M = -.014$, $SD = .300$); $d = .89$ and 1.1 for

speech and typed, respectively. There was no statistical difference between the spoken and typed conditions, $p = .466$, $d = .19$.

Subjective Evaluation of Tutorial Session. Similar to Experiment 1, there were no significant differences in participants' responses on the post-interaction questionnaire (see Table 4). As in Experiment 1, participants' impressions of AutoTutor were not affected by input modality.

INSERT TABLE 4 ABOUT HERE

Planning time, Response Time, and Verbosity. We investigated whether initial planning time, response time, and response verbosity varied across conditions. Initial planning time was measured as the time interval (in seconds) between the end of the tutor's turn and the initiation of the participant's response (detected from the F1 key press). Response time was the time interval between the initiation and completion of the participant's response (i.e., elapsed time between the F1 and F2 key presses). This measure is not process-pure as it reflects time devoted to the composition as well as production of the spoken or typed response. Response verbosity, measured as the number of words in the student's response, was separately computed for responses to main questions, pumps, hints, and student summaries.

Descriptive statistics on these six measures are presented in Table 5. Each measure was computed at the turn level and then averaged across turns to obtain an aggregate score for each participant.

INSERT TABLE 5 ABOUT HERE

Paired-sample t-tests indicated that there was a significant difference in initial planning and response times across conditions. It appears that initial planning times were approximately one-second longer when learners spoke their responses compared to when they typed them in,

t (31) = 3.49, p < .01, d = .56. A reverse pattern was obtained for response times. Here, participants responses times were substantially longer when they typed compared to when they spoke, t (31) = 14.1, p < .001, d = 2.49. The increased time associated with typed responses can potentially be attributed to the increased difficulty of typing (compared to speaking), but typing speed scores did not correlate with response times (r = -.109, p = .559). It appears that planning is more distributed in typed responses than speech over the course of a turn.

We assessed whether differences in initial planning and response times had an impact on learning gains. There were no significant correlations (p > .05) for either condition: for initial planning time, r_{spk} = .173, r_{typ} = .213; for response time, r_{spk} = -.175, r_{typ} = -.110. However, the direction and magnitude of these correlations are comparable for text and speech.

Regarding the verbosity measures, participants' responses were 14.0 words on average. This is substantially greater than the 4.4 word mean response length obtained in Experiment 1. This result substantiates that our strategy of requiring participants to articulate more content for main questions, pumps, and summaries had its desired effect. As could be expected, the most verbose responses were associated with summaries, followed by initial answers to main questions, followed by responses to hints and pumps.

There was a significant difference in verbosity across conditions for summaries t (30) = 2.16, p = .039, d = .41, but not for their responses to main questions, pumps, and hints. It appears that the summaries were approximately six-words longer when they were spoken compared to when they were typed in. However, a correlation analysis did not yield a significant relationship between verbosity of summaries and learning gains, r_{spk} = -.106, r_{typ} = .164.

Impact of Individual Differences. We assessed whether individual differences in prior knowledge, typing speed, and computer usage and knowledge were associated with differences

in learning gains across conditions. Beginning with prior knowledge, participants were assigned to either a low or high prior-knowledge group on the basis of a median split on their pretest scores. A condition × prior-knowledge ANOVA did not yield a significant interaction, $p = .663$, indicating that prior knowledge did not impact learning gains across conditions. This indicates that there were no differences in learning gains for the learners with low-prior knowledge, $M = .389$ ($SD = .308$) for speech and $M = .416$ ($SD = .371$) for typed, $p = .783$, $d = .08$. Similarly, there were no differences for learners with high prior-knowledge, $M = .170$ ($SD = .372$) for speech and $M = .263$ ($SD = .307$) for typed, $p = .435$, $d = 27$. Interestingly, although the prior knowledge main effect was not significant ($p = .065$), the data indicates that the low-knowledge learners ($M = .402$, $SD = .269$) learnt almost twice as much ($d = .69$) as their high-knowledge counterparts ($M = .217$, $SD = .270$).

We analyzed whether learning gains for the spoken and typed condition were influenced by typing speed. The mean typing speed was 27.1 words per minute ($SD = 9.86$ wpm). Participants were assigned to a slow or fast group on the basis of a median split on their typing speeds. A condition × typing speed ANOVA did not yield a significant interaction, $p = .933$.

Finally, we tested whether participants' responses on the *Computer Usage Questionnaire* interacted with learning gains for the spoken and typed condition. Instead of examining each variable independently and performing multiple statistical tests, we identified latent components in students' responses with an exploratory factor analysis. Specifically, a principal components analysis with varimax rotation and Kaiser normalization was performed on participants responses to nine out of the 11 items on the questionnaire; two items were discarded due to floor or ceiling effects (items 5 and 9 in Appendix B). Several indicators of factorability on the model indicated that the data were in fact factorable (i.e., assumptions and requirements of the factor

analysis were satisfied). In particular, (a) the included items had a correlation of at least .3 with at least one other item, which suggests reasonable factorability, (b) the Kaiser-Meyer-Olkin measure of sampling adequacy was .526, which is close to the recommended value of .60, (c) Bartlett's test of sphericity was significant; $\chi^2(36) = 63.8$, $p < .01$, (d) the diagonals of the anti-image correlation matrix were above .5 for six of the items, and above .4 for the three remaining items (which supports the inclusion of each item in the factor analysis) and (e) the commonalities were above .4, which indicates that each item shared a degree of common variance with the other items.

The analysis yielded three components with eigen values greater than 1, which together accounted for 62.6% of the variance. Component 1, which accounted for 28.1% of the variance, appears to be consistent with learners who are highly *motivated to learn* about computers (see Table 6). These learners are interested in how computers work, they enjoy learning more about computers, and are confident that they can learn from a computer tutor. In contrast to this component, Components 2 and 3 are consistent with *expertise* and *experience* with computers, respectively. Specifically, the loadings on Component 2 (20.1% of the variance) are consistent with learners who spend a lot of time working on computers and are computer users that offer more computer advice than they receive (expertise). Component 3, which explained 14.4% of the variance, is consistent with learners who have a long history of computer usage and ownership (experience). These components might appear to be similar because experience is arguably one component of expertise. Nevertheless, we considered them to be independent factors in order to be consistent with the intrinsic patterns of variability in the data as identified by the principal component analysis.

INSERT TABLE 6 ABOUT HERE

Our analyses proceeded by investigating whether learning gains for the spoken and typed conditions interacted with three components derived from the computer usage questionnaire. We addressed this question by assigning learners to a low or high group (with a median split) on each of the components and performing separate 2 × 2 (*condition*: spoken vs. type × *component*: low vs. high) ANOVAs for each component. The results indicated that the condition × component interaction was not significant for Components 2 (p = .557) and 3 (p = .737), however, there was a significant interaction for Component 1, $F(1, 30) = 6.224$, MSe = .073, p = .018, partial η^2 = .284 (see Figure 2).

Interestingly, post-hoc tests indicated that learners who scored high on Component 1 (i.e., the highly motivated learners) learned approximately twice as much when they typed (M = .468, SD = .325) compared to when they spoke their responses (M = .265, SD = .320), p = .032, d = .63. A reverse effect was observed for the less motivated learners (see Figure 2). These learners demonstrated higher proportional learning gains when speaking (M = .344, SD = .386) compared to typing (M = .207, SD = .336). Although the small sample size (only 16 learners were assigned to this group) precluded this difference from being statistically significant (p = .189), the .38 effect size this effect might be detected with a larger sample or with a more sensitive analysis.

As a more sensitive alternative to the median-split analysis, we assessed whether individual differences in motivation moderated the effect of input modality on learning. We addressed this question by performing a two-step multiple regression analysis with proportional learning gain scores as the dependent variable. Step 1 predictors were input modality (dummy coded with 1 for speech and 0 for typed) and motivation scores from Component 1 of the factor analysis. The Step 2 predictor was an interaction term for input modality and motivation (both

variables were centered prior to computing the interaction term). Of primary interest is the interaction term, which was statistically significant (p = .036), and explained an additional 7% of the variance above non-significant additive effects (Step 1 model). A simple slopes analysis revealed that standardized slopes 1 SD above and below the mean for motivation were -.24 and .11, respectively (slope at the mean was -.06). The slopes were considerably steeper 2 and 3 SD from the mean (-.43, .29, -.61, .47 for +2 SD, -2 SD, +3 SD, and -3 SD, respectively). These slopes indicate that the strongly motivated learners (SD's above the mean) learnt less when speaking compared to typing (negative slopes), while a reverse pattern was obtained for the learners lacking motivation (SD's below the mean and positive slopes). Indeed, motivation moderated the relationship between input modality and learning gains.

INSERT FIGURE 2 ABOUT HERE

We conducted a follow-up analysis to investigate *why* learners scoring high on the motivation to learn dimension learned more when typing compared to speaking their responses. The subsequent analysis focused on the 19 learners scoring above the median on the motivation factor. Our first question pertained to whether these learners preferred typing versus speaking. Paired-sample t-tests on responses to the post-interaction questionnaire indicated that there was a significant difference in participants' responses to question 7 (perceived cognitive load), but not for any of the other questions. It appears that these learners rated their cognitive load to be higher when speaking (M = 4.26, SD = 1.37) compared to typing (M = 4.68, SD = 1.53), t (18) = 2.19, p = .042, d = .29.

In a follow-up analysis we examined whether perceived cognitive load was related to learning gains. There was no relationship for the typed condition (r = .198, p = .431). However, this relationship was negative for the spoken condition, although the correlation did not reach

significance, perhaps because of the small sample size (r = -.443, p = .058). The difference between these two correlations, computed using the modified Pearson-Filon statistic (Raghunathan, Rosenthal, & Rubin, 1996), was significant ZPF (N = 19) = 2.47, p < .05. The fact that (a) the highly motivated learners reported more cognitive load when speaking and (b) cognitive load associated with spoken input was negatively correlated with learning gains in the spoken condition, might be one explanation as to why learning gains were lower when speaking compared to typing.

We also investigated whether differences in the interaction styles (i.e., initial planning time, response time, and verbosity) of these highly motivated participants could explain the patterns in learning gains. No consistent patterns emerged, so these results are not reported.

General Discussion

We contrasted text facilitation, speech facilitation, and modality equivalence hypotheses within the context of tutorial dialogues with intelligent tutoring systems. Using a Wizard of Oz experimental design, we were able to eliminate most speech recognition errors, which was a possible confound in previous experiments that investigated differences between spoken and typed student input (D'Mello, et al., in press; Litman, et al., 2006). The results of this study support some important conclusions pertaining to the impact of the input modality on learning gains and other measures of the effectiveness of tutorial dialogues. We will first discuss the theoretical implications of our findings and then address some possible applications of our results.

Implications of Findings

Our first finding was that significant learning occurred for both spoken and typed input compared to a no-intervention control condition, with mean effect sizes of .88 and .87 sigma,

respectively. The effect sizes obtained in the present experiments were within the range of effect sizes obtained by AutoTutor in over 20 experiments (Graesser, et al., 2004; Storey, Kopp, Wiemer, Chipman, & Graesser, in press; VanLehn, et al., 2007). The effect sizes in those experiments ranged from .4 to 1.5 sigma (a mean of .8), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor. The no-intervention control has been routinely used in these assessments of AutoTutor as well as other learning environments (Dodds & Fletcher, 2004; Woolf, 2009). Previous studies have also used other comparison conditions, such as a direct instruction control, but these other comparison conditions were conducted to test the impact of AutoTutor components that are unrelated and orthogonal to the contrast between spoken versus typed student input.

Another important finding was that the speech facilitation hypothesis was not supported in the present two experiments. This result is compatible with the results of a previous study with a version of AutoTutor that supported automatic speech recognition (D'Mello, et al., in press). The results of the present two experiments also align with findings reported by Litman et al. (2006) in their comparison of spoken and typed computer tutorial interactions. Taken together, these four experiments (the present two experiments with human transcriber and earlier experiments by D'Mello et al. and Litman et al. with automatic speech recognition) fail to support theories that predict the speech facilitation hypothesis. Simply changing students' input modality from text to speech does not produce higher learning gains.

Available evidence supports either the modality equivalence hypothesis or under some conditions (for motivated learners) the text facilitation hypothesis. The only exception to this generalization is Litman et al's (2006) experiment with human tutors which showed advantages for spoken over typed input. However, the Litman et al. support for spoken input with human

tutors has not been replicated and is incompatible with all of the findings available for interactions between humans and intelligent tutoring systems. Future research needs to be conducted to assess possible differences between human and computer tutors.

One potential advantage of spoken input is that the paralinguistic features of speech provide a rich trace of information related to the social dynamics, attitudes, and affective states of the learner. These additional informational channels presumably play an important role in human-human tutoring. For example, unlike current computer tutors, human tutors diagnose and respond to learners' affective states in addition to their cognitive states (Goleman, 1995; Lepper & Woolverton, 2002). Although it is difficult to compare computer tutors to human tutors directly, effective use of the paralinguistic features of speech might explain the effectiveness of human-human tutorial sessions in Litman's study (2006). Perhaps such advantages in spoken input can be replicated with computer tutors that analyse the speech channel beyond its verbal content to detect and respond to learners' affective as well as cognitive states.

Results are mixed when considering support for the text facilitation and modality equivalence hypotheses. When individual differences were not considered, there was consistent support for the modality equivalence hypothesis in the present two experiments. This hypothesis states that input modality alone has no impact on learning gains when the tutoring system is pitched for the acquisition and construction of deep knowledge. Earlier we proposed two possible explanations for modality equivalence. One was that the content is more important than the input modality, at least when it comes to learning at deeper levels of comprehension. Our second explanation for modality equivalence was that the presumed costs and benefits associated with each modality cancel each other out. Although more research is needed to decide between

these possibilities, the modality equivalence hypothesis does have some credibility in the context of interactions with computer tutors.

Nevertheless, the modality equivalence hypothesis was challenged when individual differences were analyzed in Experiment 2. It appears that learning gains were higher when the highly motivated students typed their responses, ostensibly because they experienced more cognitive load when speaking (measured via self-report scales on cognitive load) and this increased cognitive load was negatively correlated with learning. Although these results are intuitively plausible, they raise the question of why this pattern was only observed with the highly motivated learners. It is plausible that motivated learners were making more of a concentrated effort to understand the tutorial content and provide high quality answers when compared to their less-motivated counterparts. Therefore, they were more susceptible to the effects of cognitive load induced by speaking compared to typing (as the data suggest), and learning in the speech condition was consequently compromised. This trend, which merits replication in subsequent research, is suggestive of a preference for text over oral channels for the motivated students.

Applications of Findings

Although the development of automatic speech recognition technologies has been a vibrant research area for a few decades, the emergence of commercially-available affordable systems is more recent. Hence, it is important to consider the implications of our findings for ITSs and related systems that aspire to support spoken input.

There are situations when spoken input is an essential part of the interaction. Computer systems that help students learn language (Johnson & Valente, 2008; Mostow, 2008), often require spoken input. Johnson's Tactical Language and Culture Training system (Johnson &

Valente, 2008) has been used by tens of thousands of servicemen deployed in Iraq to learn the languages of the region. This system uses virtual reality and animated conversational agents to depict authentic scenarios of social interactions in a culture. Students speak their contributions and feedback is given on both their content and speech quality. Another example is Mostow's Reading Tutor (Mostow, 2008; Mostow & Aist, 2001). This tutor uses automatic speech recognition to monitor and provide feedback on spoken words when children read. Spoken input is a critical part of these learning environments and typed input will simply not suffice.

Of greater relevance are the many computer environments with natural language interaction that have the capability of supporting both spoken and typed dialogues. The results of these experiments are particularly germane for such systems. The primary suggestion is that ITSs should support typed input because all available data generally support the modality-equivalence hypothesis. Spoken-input systems might be novel, exciting, and the wave of the future, but sometimes the simplicity afforded by typed-input cannot be matched. Typed-input is cost-effective, computationally efficient, error-free (with the exception of typos and misspellings), and in most cases is equally effective in facilitating learning.

This recommendation of exclusively relying on typed input is acceptable for learners with average motivation levels, which might be all that is needed in some applications. That being said, there is also the possibility of leveraging the information gleaned in these experiments to improve learning gains by capitalizing on the differential impact of motivation on the two input modalities. The interaction we found between motivation and input modality suggests that there is some value associated with measuring the student's motivation orientation prior to the tutorial session. This information would help guide the decision of whether the interaction will be spoken or typed. Students who are highly motivated to learn would benefit most from typed

input. Less motivated learners would either be assigned speech input or would be given the opportunity to select whether they speak or type.

In summary, our discovery of an aptitude-treatment interaction with respect to motivation and modality calls for replication because of practical implications and because it theoretically challenges the null effect predicted by the modality equivalence hypothesis. It also opens the door for exploring aptitude-treatment interactions in other applications with computers that interact with humans in natural language.

References

Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*(2), 147-179.

Anderson, J., Douglass, S., & Qin, Y. (2005). How should a theory of learning and cognition inform instruction? In A. Healy (Ed.), *Experimental cognitive psychology and it's applications* (pp. 47-58). Washington, DC.: American Psychological Association.

Berlyne, D. (1978). Curiosity in learning. *Motivation and Emotion, 2*, 97-175.

Biggs, J. (1995, July). *Enhancing teaching through constructive alignment.* Paper presented at the 20th International Conference on Improving University Teaching, Hong Kong, Hong Kong.

Bransford, J., Goldman, S., & Vye, N. (1991). Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. Sternberg & L. Okagaki (Eds.), *Influences on children* (pp. 147-180). Hillsdale, NJ: Erlbaum.

Brown, A. (1988). Motivation to learn and understand - On taking charge of ones own learning. *Cognition and Instruction, 5*(4), 311-321.

Brown, A., Ellery, S., & Campione, J. (1998). Creating zones of proximal development electronically. In J. Greeno & S. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 341-367). Mahwah, NJ: Lawrence Erlbaum.

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.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Chi, M., Deleeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Chi, M., Roy, M., & Hausmann, R. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32*(2), 301-341. doi: 10.1080/03640210701863396

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, J. Levine & S. Teasely (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, D.C: American Psychological Association.

Cohen, P., Kulik, J., & Kulik, C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237-248.

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz & J. Vassileva (Eds.), *Proceedings of 8th International Conference on User Modeling* (pp. 137-147). Berlin / Heidelberg: Springer.

Corbett, A., & Anderson, J. (1994). Knowledge tracing - Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253-278.

Corbett, A., Anderson, J., Graesser, A., Koedinger, K., & VanLehn, K. (1999). Third generation computer tutors: Learn from or ignore human tutors? *Proceedings of CHI Conference on Human Factors in Computing Systems* (pp. 85 - 86). New York: ACM.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Craik, F. I. M., & Tulving, E. (1972). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

D'Mello, S., King, B., & Graesser, A. (in press). Towards spoken human-computer tutorial dialogues. *Human-Computer Interaction*.

Dodds, P., & Fletcher, J. (2004). Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia, 13*(4), 391-404.

Gergle, D., Millen, D., Kraut, R., & Fussell, S. (2004, April). *Persistence matters: making the most of chat in tightly-coupled work.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Vienna, Austria.

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 / Heidelberg: Springer.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Graesser, A., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*(4-5), 298-322. doi: 10.1080/01638530802145395

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (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: IOS Press.

Graesser, A., Ozuru, Y., & Sullins, J. (in press). What is a good question? In M. McKeown & G. Kucan (Eds.), *Threads of coherence in research on the development of reading ability*. New York: 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., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39-51.

Gratch, J., & Marsella, S. (2001, May). *Modeling emotions in the mission rehearsal exercise.* Paper presented at the 10th Conference on Computer Generated Forces and Behavioral Representation, Piscataway, NJ.

Johnson, W., & Valente, L. (2008, July). *Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures.* Paper presented at the 20th National Conference on Innovative Applications of Artificial Intelligence, Chicago, Illinois.

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.

Lepper, M., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135-158). Orlando, FL: Academic Press.

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.

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*(2), 145-170.

Litman, D., & Silliman, S. (2004, May). *ITSPOKE: An intelligent tutoring spoken dialogue system.* Paper presented at the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics, Boston, MA.

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

Mayer, R., Sobko, K., & Mautone, P. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*(2), 419-425. doi: 10.1037/0022-0663.95.2.419

Moshman, D. (1982). Exogenous, endogenous, and dialectical constructivism. *Developmental Review, 2*(4), 371-384.

Mostow, J. (2008). Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods. In C. K. Kinzer & L. Verhoeven (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments Through Technology* (pp. 117-148). Mahwah, NJ: Erlbaum.

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.

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

Palinscar, A., & Brown, A. (1984). Reciprocal teaching and comprehension fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117-175.

. PC Usage Questionnaire. (2010) Retrieved Jan 23, 2010, 2010, from http://www.busreslab.com/consult/compuse.htm

Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R. H., & Raymond, P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*(3), 531-549.

Person, N., & Graesser, A. (2002). Human or computer? AutoTutor, in a Bystander Turing Test. In S. Cerri, G. G. & P. F. (Eds.), *Proceedings of the 6th International Conference on Intelligent Tutoring Systems* (pp. 821-830). Berlin / Heidelberg: Springer.

Piaget, J. (1952). *The origins of intelligence*. New York: International University Press.

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E. O., & Peters, S. (2004). Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. In J. Lester, R. Vicari & F. Paraguacu (Eds.), *Proceedings of Seventh International Conference on Intelligent Tutoring Systems* (pp. 390-400). Berlin / Heidelberg: Springer.

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

Quinlan, T. (2004). Speech recognition technology and students with writing difficulties: Improving fluency. *Journal of Educational Psychology, 96*(2), 337-346. doi: 10.1037/0022-0663.96.2.337

Raghunathan, T. E., Rosenthal, R., & Rubin, D. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1*(2), 178-183.

Rogoff, B. (1990). *Apprenticeship in thinking*. New York: Oxford University Press.

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*(1), 23-52.

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

Sweller, J. (1988). Cognitive load during problem-solving - Effects on learning. *Cognitive Science, 12*(2), 257-285.

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

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research, 64*, 37-54.

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.

VanLehn, K., Jones, R., & Chi, M. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2*(1), 1-59.

VanLehn, K., Jordan, P., Rose, C., Bhembe, D., Bottner, M., & A., G. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G.

Gouarderes & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring* (pp. 158-167). Berlin: Springer-Verlag.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

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). NJ: Erlbaum.

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

Table 1

Hypothesized Benefits of Typed and Spoken Input

Property	Typed	Speech	Benefits
Persistence of text on screen	+		Enhanced encoding due to increased perceptual processing
Persistence of text on screen	+		Better quality of responses because the responses can be inspected and revised
Text can be offloaded to an external record (i.e., input box)	+		Planning and response composition can be intertwined
Text can be offloaded to an external record	+		Reduces cognitive load especially during composition of lengthy responses
Effortless production		+	Reduces cognitive load associated with response construction
Effortless production		+	Facilitates longer responses and there is evidence that response verbosity positively correlates with learning
Simultaneous response production while attending to other areas of screen		+	Frees up (visual) attentional resources when responses involve the inspection of images
Presence of social cues		+	Deeper processing according to social agency theory

Notes. The same property can have multiple benefits as is the case with "persistence of text on screen". Also note that "persistence of text on screen" and "text can be offloaded to an external record" are in fact the same property. They have been listed separately as they have conceptually different effects.

Table 2

Sample Conversation between AutoTutor and a Relatively Verbose Student

Text	Dialogue Move
[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	Contribution
[AutoTutor] Anything else?	Pump for information
[Student] This file contains the operating system	Contribution
[AutoTutor] What about storage?	Hint
[Student] The operating system is then read into RAM and activated.	Contribution
[AutoTutor] Right. The CPU executes instructions in ROM, which helps the operating system to start.	After giving positive feedback, there is an assertion

Table 3

Learner Satisfaction for Spoken and Typed Conditions for Experiment 1

	Speech		Typed	
Item	*M*	*SD*	*M*	*SD*
(Q1) I enjoyed interacting with AutoTutor	3.91	1.44	3.52	1.65
(Q2) Interaction comparable to human tutor	3.44	1.31	3.57	1.47
(Q3) AutoTutor did *not* understand what I said	3.00	1.24	3.44	1.24
(Q4) I felt engaged during the tutoring session	3.70	1.46	3.57	1.59
(Q5) I felt that AutoTutor was difficult to use and work with	2.74	1.31	3.22	1.68
(Q6) I felt that I learned new information from AutoTutor	4.35	1.40	4.22	1.45
(Q7) Difficulty of learning Computer Literacy	3.74	1.60	4.09	1.98
(Q8) Difficulty of learning new information	3.30	1.61	3.57	1.75

Note. A six-point scale from (1) *strongly disagree* to (6) *strongly agree* was used for questions 1 to 6. A seven-point scale from (1) *very easy* to (7) *very hard* was used for questions 7 and 8.

Table 4

Learner Satisfaction for Spoken and Typed Conditions for Experiment 2

	Speech		Typed	
Item	*M*	*SD*	*M*	*SD*
(Q1) I enjoyed interacting with AutoTutor	2.89	1.43	2.97	1.52
(Q2) Interaction comparable to human tutor	2.47	1.54	2.29	1.38
(Q3) AutoTutor did *not* understand what I said	3.43	1.56	3.37	1.65
(Q4) I felt engaged during the tutoring session	3.34	1.55	3.46	1.65
(Q5) I felt that AutoTutor was difficult to use and work with	3.54	1.54	3.49	1.65
(Q6) I felt that I learned new information from AutoTutor	4.83	0.80	4.86	0.79
(Q7) Difficulty of learning Computer Literacy	4.82	1.38	4.62	1.35
(Q8) Difficulty of learning new information	4.47	1.38	4.47	1.64

Note. A six-point scale from (1) *strongly disagree* to (6) *strongly agree* was used for questions 1 to 6. A seven-point scale from (1) *very easy* to (7) *very hard* was used for questions 7 and 8.

Table 5

Initial Planning Time, Response Time, and Verbosity for Experiment 2

	Speech		Typed	
Measure	*M*	*SD*	*M*	*SD*
Initial planning time (seconds)	4.96	2.03	4.05	1.11
Response time (seconds)	7.22	2.89	16.3	4.27
No. words in response to main questions	12.3	5.09	10.7	5.66
No. words in response to pumps	10.3	10.6	8.77	5.67
No. words in response to hints	6.18	3.30	6.37	2.37
No. words in response to summaries	30.7	13.1	25.3	13.1

Table 6

Component Loadings for Computer Usage and Knowledge Questionnaire

		Component		
Item		1	2	3
I10	Interest in how computers work	.765		
I6	Find computers easy to use	.713		.500
I11	Confidence can learn from computer tutor	.706		
I7	Enjoy learning new computer software	.657		
I4	Level of understanding of computer usage		.765	
I8	Give computer advice to others		.758	
I1	Hours per week of computer usage		.727	
I12	Number of years of computer usage			.885
I12	Number of years of computer ownership			.673

Note. Items sorted by size and values < .4 are suppressed. Identifiers (I1, I2) correspond to items on the questionnaire (see Appendix B).

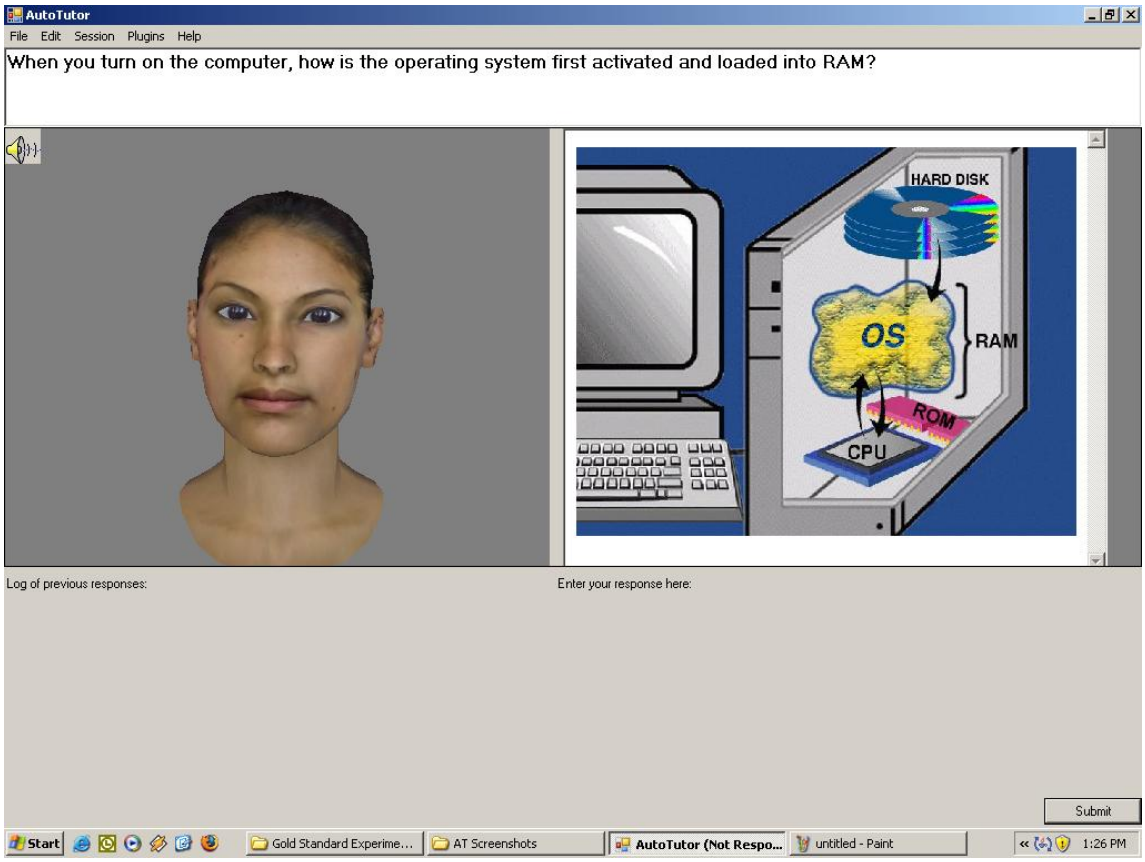

Figure 1. The AutoTutor Interface

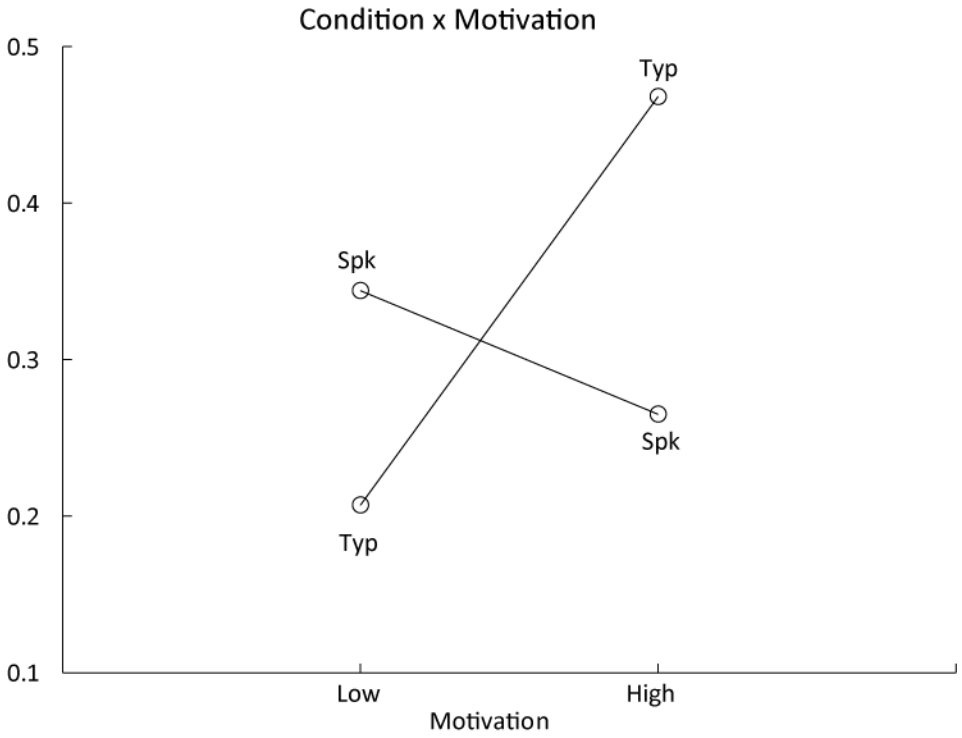

Figure 2. Condition x Motivation Interaction for Experiment 2. Y axis represent proportional learning gains.

Appendix A. Example Questions from Knowledge Tests

1. *If you install a sound card, why does your computer perform better? [Hardware Question]*
    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
    d. Because there will be fewer bottlenecks when multitasking

2. How does the computer assure that other stored information is not overwritten when a save command is given? [Operating Systems Question]
    a. The application is able to write the document to unused space on the hard drive
    b. The operating system uses one of the utility programs designed for that purpose
    c. The operating system communicates with the CPU
    d. The operating system communicates with RAM

3. When you get an e-mail message across the Web, what computers does it travel through? [Internet Question]
    a. Your computer, your server, and the computer at the address you get it from
    b. Your computer, your server, the server for the computer you get it from, and the computer at the address you get it from
    c. Your computer, your server, the NORAD computer, the server for the computer you get it from, and the computer at the address you get it from
    d. Many computers

Appendix B. Computer Usage and Knowledge Questionnaire

1. *About how many hours per week do you use a computer?*

   (a) 0 hrs (b) 1 to 5 hrs (c) 5 to 10 hrs (d) 11 to 20 hrs (e) 21 to 30 hrs (f) 30 or more hrs

2. *About how long have you been using a computer?*

   (a) < 1 year (b) 1 to 2 years (c) 2 to 5 years (d) 5 to 10 years (e) > 10 years

3. *For how long have you owned your own computer*

   (a) I don't own one (b) < 1 yr (c) 1-2 yrs. (d) 2-5 yrs. (e) 5-10 yrs. (f) > 10 yrs.

4. *Which statement below best describes your level of understanding about how to use a personal computer/PC*

   (a) I don't know how to turn on a personal computer. (b) I can turn on a PC, but I have trouble with almost everything else. (c) I know the basics of how to use my PC, but not much more (d) I understand how to use most of my software, and have little trouble learning new software. (e) I completely understand my PC software and hardware.

5. *I wish computers had never been invented*

   6-point scale from strongly disagree to strongly agree (same scale for items 6-8)

6. *I find computers extremely easy to use*

7. *I really enjoy learning new computer software*

8. *I give more computer advice to other people than I receive*

9. *Have you ever interacted with a computer through speech (yes or no)*

10. *How interested are you about how computers work (software and hardware)?*

    6-point scale from very disinterested to very interested

11. *How confident are you that you can learn from a computer tutor?*

    6-point scale from not at all confident to very confident