Comparing Human-Human to Human-Computer Tutorial Dialogue

Natalie B. Steinhauser (Natalie.Steinhauser@navy.mil) &
Gwendolyn E. Campbell (Gwendolyn.Campbell@navy.mil)
Naval Air Warfare Center Training Systems Division, Code 4.6.5.1
12350 Research Parkway, Orlando, FL 32826-3275

Katherine M. Harrison (Katherine.M.Harrison.ctr@navy.mil)
Kaegan Corporation
12000 Research Parkway, Orlando, FL 32826-2944

Leanne S. Taylor (Leanne.Taylor.ctr@navy.mil)
University of Central Florida
4000 Central Florida Blvd. Orlando, FL 32816

Myroslava O. Dzikovska (M.Dzikovska@ed.ac.uk) & Johanna D. Moore (J.Moore@ed.ac.uk)
Human Communication Research Centre, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom

Abstract

Intelligent Tutoring Systems are often modeled after human tutors; however, the effectiveness of this strategy is yet to be determined. Research on media interactions suggests that behaviors with humans are similar to those with computers. Intelligent Tutoring System studies have said the opposite. In this study we compared a human-human and a human-computer tutoring system in terms of metacognitive, social, and nonsense statements to dig deeper into these interactions. We discovered that the interactions were quite different between human-human and human-computer tutoring. With a human, participants expressed more positive metacognitive statements and social statements. When interacting with a computer tutor, students were more likely to make negative metacognitive statements and social statements. In addition, the interpretation of these results differed between the two corpora. In human-human tutoring, the more often a participant made positive metacognitive statements, the worse their learning gain. Their social dialogue had no impact on learning gain. In human-computer tutoring, the more negative and positive metacognitive statements and the more negative social statements they gave the worse their learning gain. It is clear from this study that students do not act the same with a human tutor as they do with a computer tutor. Therefore, designers of ITS systems should not just blindly model their systems after human tutors. The differences in human and computer interactions should also be considered.

Keywords: Human Computer Interaction (HCI), Intelligent Tutoring Systems (ITS), Metacognition, Social dialogue, Tutorial dialogue

Introduction

Over the years, Intelligent Tutoring Systems (ITSs) have become popular learning and teaching tools. Thus, their design is becoming more sophisticated. One approach to creating ITSs is to model them after a human tutor because human tutoring has been said to be the most effective form of teaching (Bloom, 1984). However, it has not yet been determined that this is a good strategy. Two unresolved questions are whether you will find the same kinds of dialogue when a student is interacting with a human and a computer tutor (ITS) and whether those types of dialogue can be interpreted in the same way with regards to the learning that is occurring.

Research on media interactions has stated that people interact socially and naturally with media (to include computers) as they do humans (Reeves & Nass, 1996). The researchers suggest that people follow rules of social relationships when interacting with media and that this occurs naturally and unconsciously. For example, media has been shown to induce emotions such as frustration and politeness.

Similarly, studies examining interactions with virtual humans have shown that people react in the same manner to these entities as they do with other humans (Zanbaka, Ulinski, Goolkasian, & Hodges, 2004; Pertaub, Slater, & Barker, 2002). While being observed by a crowd of virtual agents, people showed nervousness just as they did with a human audience. Women also show social inhibition effects with virtual agents like they do with humans.

In contrast, more recent research using ITSs has shown that students do not behave the same with computers as they do with humans, as evident in their dialogue acts. When students were conversing with a computer, but believed they were conversing with a human, they used more words and conversed longer than did students who were told they were talking to a computer (Schechtman & Horowitz, 2003). In addition, students provided more explanations and longer turns when they believed they were talking to a human versus a computer, even though they were talking to a computer in both cases (Rosé & Torrey, 2005).

Therefore, results as to how people respond to computers and computer entities, in comparison to humans, are mixed. While previous ITS studies have looked at the content based dialogue (dialogue relevant to the lesson material), we took

a broader perspective and considered other dialogue categories, such as metacognition, because they have also been shown to predict learning gain (Campbell et al., 2009). We examined and compared a human-human and a human-computer tutorial dialogue corpus. We categorized five types of dialogue found in these corpora. Most of the dialogue was related to the content of the lessons. The other four categories of dialogue that were present were management (discussing the flow of the lesson), metacognition (describing one's understanding), social (chit-chat and signs of frustration), and nonsense words (random sequences of letters). For this comparison, we will focus on metacognition, social dialogue, and nonsense words because these are the categories where research hasn't yet explored and, we believe, will also differ in regards to the interactions.

Method

To explore our research questions we conducted a human-human and a human-computer study. The two corpora were then analyzed and compared in terms of their dialogue.

Human-Human Tutoring Study

Data collection environment

A curriculum incorporating lessons on basic electricity and electronics was constructed. The curriculum covered topics including open and closed paths, voltage reading between components and positive and negative terminals, series and parallel configurations, and finding faults in a circuit with a multimeter. These basic concepts were taught in a computer-based learning environment within a single session lasting approximately four hours[1].

Figure 1 shows a screenshot of the learning environment that the participants interacted with during the study. The screen was divided into three sections. The top left-hand section displayed the core lesson material in slide form, including educational text, activities, and discussion questions. The participants were able to move through the lesson slides at their own pace. The top right-hand section provided participants with a circuit simulator, which allowed them to construct and manipulate circuits as a supplement to the material in the slides. The bottom section was the chat window where the participants and tutor conversed by typing.

The tutor and student were located in the same room, but were separated by a divider. The tutor had the ability to observe the student's learning environment and interact with the student through a computer screen and chat window. The tutor gave feedback, technical assistance, and/or encouragement that he or she considered appropriate. Participants directed their answers, comments, and/or questions to the tutor throughout the curriculum.

[1] Note that there was a second session, covering additional topics, but it will not be addressed further in this paper.

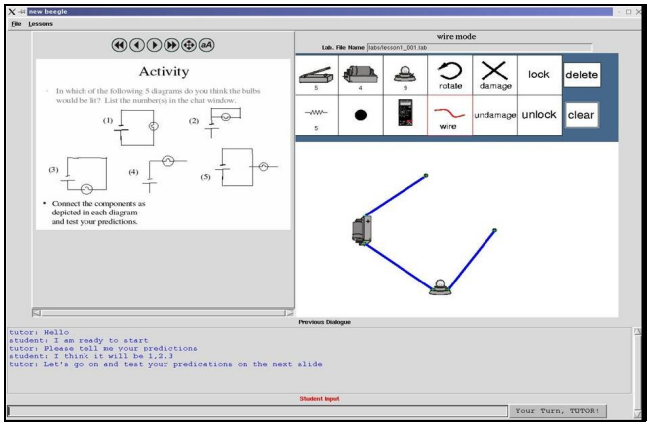

Figure 1. Participant screen for human-human tutoring.

Procedure

After completing informed consent paperwork, participants filled out a demographic questionnaire and took a pre-test consisting of 38 multiple choice questions. The participants were then introduced to their tutor and given a brief demonstration of how to operate the learning environment. The students spent the majority of the experimental session working through the lesson material and building circuits. At the conclusion of the experiment, participants completed a post-test, which included 21 multiple choice questions, and a reaction questionnaire. They were then debriefed and excused.

Corpus

The corpus of the human-human study was comprised of dialogues from each of the thirty participants distributed across three experienced tutors. The average age of the participants was 22.4 years (*SD* = 5.0) and exactly half of them were female. The corpus of this study includes 8,085 dialogue turns taken by the student and tutor and 56,133 tokens (words and punctuation).

Human-Computer Tutoring Study

Data collection environment

As much as possible, the same curriculum as the human-human study was used in the BEETLE II computer tutoring system (Dzikovska et al., 2010). Small changes were made to the curriculum so that the computer would be able to understand student responses (e.g., multi-part questions were simplified into single questions). The computer tutor (ITS) was created to implement the effective tutorial strategies used in our human-human corpus (e.g., hints). The ITS understood and responded to content (by providing feedback) and negative metacognitive statements (by giving a hint) made by a student, but not to the other types of dialogue (management, social, and nonsense). The responses and feedback given by the ITS was modeled after the human tutors from the previous corpus. The ITS used a friendly and encouraging tone similar to the human tutor. In

fact, in most cases, the ITS used identical phrasing for its comments to the student.

A screenshot of the learning environment is shown in Figure 2. The learning environment was similar to that of the human-human environment. The screen was divided into three sections. The upper left-hand section had the same function as the previous study; however the navigation buttons were slightly different. The right-hand section was the chat window where the participants and tutor interacted through typing. The lower-left section included the circuit simulator, which had the same purpose as the previous study, although the tools used to build circuits had a different display interface.

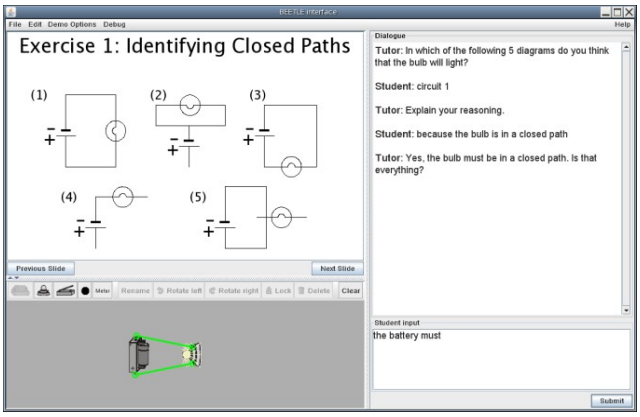

Figure 2. Participant screen for the BEETLE2 ITS.

Procedure

The procedure for the human-computer study was essentially the same as the human-human study with a few exceptions. The pre-test consisted of 22 multiple choice questions and the post-test consisted of 21 multiple choice questions. The human-computer pre-test had fewer questions because we removed questions associated with material from the second session of the human-human study, as mentioned earlier. In addition, instead of a reaction questionnaire at the conclusion of the study, participants were given a usability and satisfaction questionnaire.

Corpus

The human-computer corpus consists of dialogues from each of the forty-one participants in the study. The average age of the participants was 20.8 years (*SD* = 3.30) and there were almost twice as many females as males. The corpus includes an estimated 34,900 total dialogue turns taken by the student and tutor and an estimated 398,410 total tokens. There were many more dialogue turns and total tokens in the human-computer study because the computer asked the questions in this study (versus them being presented on the lesson slides in the previous study). In addition, more questions were presented in this study because, as stated earlier, multi-part questions were simplified into individual questions.

Coding

For the human-human data, two independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, and social dialogue statements with perfect reliability (kappa = 1.00). In addition, raters were able to differentiate between positive and negative metacognitive statements made by the student with high inter-rater reliability (kappa = 0.99).

For the human-computer data, four independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, social dialogue, and nonsensical statements with high reliability (kappa = 0.88). In addition, raters were able to differentiate between positive and negative metacognitive statements made by the student with high inter-rater reliability (kappa = 0.96).

A summary of the codes used in this study are presented in Table 1.

Content statements were described as comments including domain concepts that pertained to the lesson material. Answering a question fit into the content category (e.g., "The battery and the bulb in diagram 1", "1.5 volts", etc.).

Management consisted of dialogue that dealt with the flow of the lesson but does not contain information relevant to the lesson topics (e.g., "I give up", acknowledging the tutor's instructions to continue by saying, "OK", etc.).

Metacognitive statements were defined as statements that contained the student's feeling about his or her understanding, but did not include domain content. Metacognitive statements were further classified as positive or negative. Positive metacognitive statements were defined as statements that expressed understanding (e.g., "I get it", "I understand", etc.), whereas negative metacognitive statements expressed confusion (e.g., "I don't understand", "Give me a hint", etc.).

Social dialogue includes positive and negative statements. Positive social dialogue was defined as statements that included humor, rapport, chit-chat, and saving face. Examples included "Ha-ha", "Hi, how are you?", etc. Negative social statements included expressions of frustration, explicit refusals to cooperate, and even offensive statements. Examples included "Because I said so", "No", "You're stupid", expletives, etc.

Nonsense was classified as statements that were made up of random letters or numbers that are not content related (e.g., "ufghp", "3i9f", etc.). Nonsense did not occur in the human-human dialogue; therefore it was not coded in those transcripts.

Since we wish to look beyond just the content dialogue, we will focus on metacognition, social dialogue, and nonsense words in our results. Management was left out of the analyses because it was not very prevalent in the computer tutoring data and, when it was, it was ignored by the tutor. Also, it was not a relevant predictor of learning gain with the human tutor.

Table 1. Coding summary

Code	Definition	Example
Content	Statements including domain concepts that pertain to the lesson	"There is a battery and bulb in circuit 1." "1.5 volts."
Management	Dialogue that does not contain information relevant to the lesson material, but deals with the flow of the lesson	"I give up." "O.k." Acknowledging the tutor's instructions to continue
Metacognition	Statements containing the student's feelings about his or her understanding, but does not include domain concepts	Metacognitive statements can be positive or negative.
Positive	Statements that express understanding	"I get it." "I understand." "Oh, o.k."
Negative	Statements that express confusion	"I don't know." "I don't understand."
Social Dialogue	Dialogue that is not related to the content of the lessons and serves as motivation, encouragement, humor, frustration outlets, etc.	Social statements can be positive or negative.
Positive	Statements that include humor, rapport, chit-chat, or saving face	"Ha-ha" "Hi, how are you doing?"
Negative	Statements that include frustration, refusal to cooperate with the system, or offending the system	"Because I said so." "No." "You're stupid." Expletives
Nonsense	Random sequences of letters or numbers that do not pertain to the lesson material	"oidhf" "dsfafadgdfh"

Results

Learning Gain

Pre- and post-test scores were calculated in terms of percentage correct. A learning gain score was then calculated for each participant using the formula: (post-test score – pre-test score)/(1- pre-test score).

Metacognitive Statements

Students made metacognitive statements in both studies, regardless of whether the tutor was a human or a computer; however, the relative frequencies of positive and negative metacognitive statements depended upon the type of tutor. Specifically, students talking to a human tutor made significantly more positive metacognitive statements (M = 12.9, SD = 8.3) than negative metacognitive statements (M = 1.8, SD = 2.0), $t(28) = 7.16$, $p < 0.001$. Students talking to a computer tutor, on the other hand, made significantly more negative metacognitive statements (M = 3.8, SD = 5.5) than positive metacognitive statements (M = 0.2, SD = 0.5), $t(39) = -4.21$, $p < 0.001$.

The implications of the presence of metacognitive statements also varied depending upon the type of tutor. For students interacting with a human tutor, the amount of positive metacognitive dialogue, but not negative metacognitive dialogue, was significantly negatively correlated with learning gains; $r = -0.543$, $p = .002$ and $r = -0.210$, $p = 0.266$, respectively. However, for students interacting with the computer tutor, the frequency of both types of statements were negatively correlated with learning gains (positive statements: $r = -0.419$, $p = 0.006$; negative statements: $r = -0.537$, $p < .001$).

Social Statements

While students made social statements with both types of tutors, students interacting with a human tutor made exclusively positive social statements and students interacting with the computer tutor made exclusively negative social statements. On average, students interacting with a human tutor typed 37.5 positive social words to their tutor (SD = 52.3) and students interacting with the computer tutor typed 8.5 negative social words (SD = 20).

Interestingly, the amount of social dialogue with human tutors was unrelated to student learning gains, $r = -0.211$, $p = 0.262$, but the amount of social dialogue that the student produced when interacting with the computer tutor was negatively correlated with learning gains, $r = -0.372$, $p = .017$.

Nonsense

Finally, as mentioned earlier, students spontaneously exhibited a novel type of "utterance" when interacting with the computer tutor – nonsensical sequences of letters and/or numbers. On average, the students submitted nonsense to the computer tutor 1.7 times. There was quite a bit of variability across students in their likelihood of exhibiting this behavior, with a standard deviation of 5.1. This behavior was not statistically related to learning gains, $r =$

-0.073, $p = 0.651$. However, not surprisingly, the frequency of this behavior was significantly negatively correlated to the students' report of satisfaction with the computer tutor, $r = -0.33$, $p = 0.035$.

Discussion

As stated before, ITSs are often modeled after human tutors, but it is uncertain whether these interactions are similar and can be interpreted in the same manner. In fact, we found that students did not respond similarly to the computer tutor as they did with the human tutor. In both corpora, student's dialogue included metacognitive statements, but the nature of those statements was very different. With a human tutor the statements were mostly positive acknowledgements, whereas with the computer they were negative statements expressing confusion.

Social dialogue differed drastically as well. With a human the social dialogue was all positive and served the purpose of creating rapport. With the computer, the social dialogue was all negative and was concerned more with showing frustration with the system. Nonsense did not occur in the human corpus at all. This was a new category that occurred in the computer corpus only.

The human-human and human-computer dialogues also differ in their interpretations, specifically in the metacognition and social categories. In the human-human corpus, metacognition was a negative predictor of learning gain only when it consisted of positive statements. The more frequently students said things like "I get it" the worse they did. In the human-computer corpus, both types of metacognitive statements (positive and negative) were a bad sign, though they rarely gave positive metacognitive statements.

Social interactions also differ in their interpretation. With human-human tutoring, social dialogue was not related to learning gain, whereas in human-computer tutoring it was negatively correlated with learning gain. The social statements made in the ITS environment were all negative, reflecting the participant's frustration. Thus, expressing frustration through social dialogue was a good indicator that the student was struggling with the content.

These results indicate that interactions and interpretations may indeed be different between human-human and human-computer tutoring. They also suggest that perhaps human tutors are able to handle negative metacognitive statements like "I don't get it" more effectively than our computer tutor, since negative metacognition was not negatively correlated with learning gain in the human-human corpus.

Overall, it appears that politeness may be playing a role in human-human interactions, but is put aside in human-computer interactions. When conversing with another human, participants positively acknowledged what their tutor said and participated in rapport building with chit-chat. This seems to be driven by a need to be polite and courteous to the tutor, but wasn't a good indicator of what was really going on as far as learning was concerned. Based on the results, you may not be able to really trust a student who says "I understand" when they are interacting with a human because it is unclear if they really understand or if they are just being polite.

On the other hand, when interacting with a computer tutor, participants appear to be more honest in terms of their negative statements. If they show signs of confusion or frustration, they really seem to be indicating that they are struggling with the lessons. Such signs can be interpreted as more accurate indicators that additional remediation is needed. The rules of politeness are ignored and the true story seems to emerge.

From this study we found that students will not necessarily act the same with a computer tutor as they do with a human tutor. This suggests that designing an ITS to try to mimic a human tutor may not be the best strategy. The differences in interactions should also be considered. For example, positive social statements were not related to learning gains, so they do not necessarily need to be supported in an ITS; however, negative social and nonsense statements were negatively correlated with learning gains in the ITS and should be addressed. Perhaps additional help should be given or students should be offered a break when these forms of dialogue occur. All forms of metacognition impacted learning gain in the human-computer corpus, thus they should all be addressed in the ITS. Possibly giving additional remediation to students who make metacognitive statements could be helpful.

While modeling a human tutor may be a reasonable first step in the design of an ITS, the design cannot stop there. The ITS needs to be evaluated and tested with users to determine its effectiveness. Tweaks to the system should be made according to the ITS evaluation, like the ones suggested above, for each individual system and curriculum.

In this study we tried to model the human tutor as much as possible, but were limited by the current technological capabilities in computational natural language processing. Further advancements and improvements to the system's capabilities might yield different results. Additionally, these comparisons should be replicated in other domains and other curriculums to see how results compare. It would also be interesting to compare human-human and human-computer tutoring with spoken dialogue to see if the results would hold since tutoring is commonly done in spoken form.

Acknowledgments

We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez, three former Research Associates who worked on this project, Leslie Butler, Lisa Durrance, and Cheryl Johnson, and two additional team members, Elaine Farrow and Charles Callaway for their contributions to this effort.

References

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

Campbell, G.E., Steinhauser, N.B., Dzikovska, M.O., Moore, J.D., Callaway, C.B., & Farrow, E. (2009, July). Metacognitive awareness versus linguistic politeness:

Expressions of confusions in tutorial dialogues. Poster presented at the *31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

Dzikovska, M.O., Moore, J.D., Steinhauser, N.B., Campbell, G.E., Farrow, E., & Callaway, C.B. (2010). Beetle II: a system for tutoring and computational linguistic experimentation. *In Proceedings of ACL-2010 demo session.*

Pertaub, D., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audiences. *Presence: Teleoperators and Virtual Environments*, *11(1),* 68-78.

Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, UK: Cambridge University Press.

Rosé, C.P. & Torrey, C. (2005). Interactivity and expectation: eliciting learning oriented behaviour with tutorial dialogue systems. *Proceedings of INTERACT* (pp. 323-336).

Schechtman, N. & Horowitz, L.M. (2003). Media inequality in conversation: how people behave differently when interacting with computers and people, In *Digital Sociability*, *5(1),* 281-288.

Zanbaka, C., Ulinski, A., Goolkasian, P., & Hodges, L.F. (2004). Effects of virtual human presence on task performance. *Paper presented at the International Conference on Artificial Reality & Telepresence*. Retrieved from http://www.vrsj.org/ic-at/papers/2004/S4-1.pdf .