Conversational Agents and Learning Outcomes: An Experimental Investigation

Bob Heller
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
bobh@athabascau.ca

Mike Procter
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
mikeprocter@shaw.ca

Abstract: An experimental approach was used to compare two types of web interfaces (text-based vs. conversation-based) to content on the life and theories of Jean Piaget. The content in each interface was identical with the exception of third- vs. first-person references. Fifty-nine students in psychology first completed a pretest of Piagetian knowledge and then were randomly assigned to one of the two interfaces. After 20 minutes of review/conversation, students completed a 35-item exam designed to measure knowledge retention and a questionnaire to measure their perceptions of the assigned interface. Contrary to expectations, the text-based interface was rated significantly higher on measures of enjoyment and utility and led to better learning outcomes in comparison to the conversational agent. Altogether, the findings indicate that the use of conversational agents in distance education needs to be carefully matched to the learning goals and outcomes.

The use of conversational agents in distance education falls under the broader category of pedagogical agents, or the design of computer software that is autonomous, interactive, anthropomorphized, and directed towards educational goals and outcomes The design of pedagogical agents is guided by a number of theoretical frameworks drawn from different disciplines. For example, the work of Graesser and colleagues on AutoTutor, a conversational intelligent tutor system (see Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Tutoring Research Group 1999), is based largely on social constructionist theories of learning (e.g. Fosnot, 1996) that emphasize collaboration and conversation as a natural and effective means of knowledge construction and elaboration. Another approach reflected in the research by Cassell and colleagues on Embodied Conversational Agents (ECA) highlights the primacy of conversation as a uniquely human skill learned early and effortlessly in life and one that is essential to social interaction (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2000). Regardless of the approach pedagogical agents have been identified as an important research field at the "juncture of HCI and educational psychology" with the potential to improve learner interest, retention, and knowledge transfer (Hadwin, Winne, & Nesbit, 2005).

According to Johnson, Rickel, and Lester (2000), pedagogical agents have the potential to increase the bandwidth of communication between students and computers and increase the computer's ability to engage and motivate students. It is further assumed that these two features, traditionally problematic in distance education, ultimately improve learning outcomes and learning experiences. Lester et al. (1997) refer to these predictions as the 'persona effect' based on their findings that the mere presence of an animated agent had a strong positive effect on learner perceptions of the learning experience. Based on the persona effect, Heller, Procter, Mah, Jewell, & Cheung (2005) argued that historical figures were ideal applications for conversational agents because of the intrinsic appeal of fame and famous people. Chatting with a historical figure via the internet was hypothesized to be intrinsically more interesting than the same information presented in a standard third person text format. If engagement and motivation underlie the persona effect, then a famous historical figure should provide a strong test.

As a proof of concept, Heller et al. (2005) developed Freudbot, a conversational agent programmed to chat in the first person about Freudian theory, concepts and biographical events. In addition to the text windows used to display

user input and agent output, a photo image of Sigmund Freud was displayed on the side and partially animated to appear as though the lips were moving as new output was presented and also with smoke occasionally wafting from a depicted cigar. Freudbot was developed using Artificial Intelligence Markup Language, an XML-based open-source programming language that supports ALICE (http://www.alicebot.org), an award winning conversational agent or chatbot and progenitor of thousands of other chatbots as hosted on Pandorabots (www.pandorabots.com).

An important feature in conversational exchange and learning is the evocation of stories or narratives (Heo, 2004). To explore this natural feature of communicative exchange, Freudbot was designed to represent topical content using a narrative structure similar in function to the story grammar approach of Thorndyke (1977) and Rummelhart (1975). In total, close to 150 autobiographical and conceptual narratives were created that could be evoked by a user using certain keywords and phrases. If a user asked about a particular concept or episode, Freudbot was designed to provide an answer with implicatives that would invite the user to request more information using conversational directives (e.g. go on, tell me more, is that all, why is that, etc.), a feature consistent with the conversational rules related to turn-taking. The learner effectively controls the way in which a story can be told by switching to other stories or entering into specific parts of a story. Freudbot also had the capacity to deviate from the story and then come back.

In addition, we also developed agent strategies loosely consistent with Speech Act theory that direct the user to ask about content within Freudbot's repertoire. In cases in which no input was recognized, Freudbot would default to one of several conditional strategies: ask for clarification, suggest a new topic for discussion, indicate that he had no response, or ask the user for a suggested topic. Another type of conversational pickup strategy was to ask a question and then redirect to a known topic. For example, Freudbot could ask a question like "Are you happy?" and regardless of how the user responded, Freudbot would respond with a leading statement like "That reminds me of the pleasure principle." This feature was designed to lead the user back to a discussion of Freudian topics.

The findings with Freudbot indicated a neutral endorsement overall but participants did note the potential of conversational agents and recognized that improvements could be made. As a learning activity, chatting with a famous personality is natural and intuitive and borne out by the observations that 90% of the participant's chat behaviour was classified as on-task and that the proportion of on-task behaviour was positively correlated to overall evaluations. When participants were separated in groups based on whether they would chat again if given an opportunity, almost 70% of the sample indicated yes. Not surprisingly, those that would chat again rated the performance of Freudbot significantly higher on all 7 measures than did those who would not.

Famous personality applications were also rated significantly more highly than three other course-related applications (Course Administration, Course Content, Chatroom Bots). In the words of one participant; "It was pretty cool the way it felt like I was actually interacting with Freud... he's deceased though, yeah, but the picture, the fast answers... made me pay attention to the answers a lot more than if I had been simply reading a text written by someone else. Plus it was cool to feel like I could voice my own opinion with the most well-known psychoanalyst of all time."

Although the findings with Freudbot were useful, there is still no direct evidence on the utility of a historical figure conversational agent in comparison to a non-conversational presentation. Nor was there evidence on whether positive learning outcomes are facilitated by interacting with conversational agents. A review of pedagogical agent research indicated that learner outcomes following agent interactions are not consistently reported in the literature. Lester et al. (1997) were the first to report an increase in problem solving ability following students' interactions with a fully animated APA compared to less animated agents and other researchers have reported similar findings (e.g. Mayer, Dow, & Mayer, 2003; Moreno, Mayer, Spires, & Lester, 2001; Moundridou & Virvou, 2002). In contrast, other researchers have found no differences between groups on measures of learner outcomes (André, Rist, & Müller, 1999; Atkinson, 2002; Craig, Driscoll, Gholson, 2004). Much of this literature is difficult to interpret because of weak experimental designs (Clark & Choi, 2005; Dehn & van Mulken, 2000). Moreover, Gulz and Haake (2006) also raise concerns about a number of visual appearance variables that may be systematically affect results. In our own study, the choice of Sigmund Freud may have led to spurious findings given the controversy surrounding his career.

In order to address these shortcomings, a second historical figure application was developed using the organizational structure and control mechanisms of Freudbot but designed to emulate Jean Piaget. The same content used to create

Piagetbot was also used to create a set of text-based web pages. The only difference between interfaces was the use of the first person in Piagetbot vs. third person in the text-based web page. Following the recommendations of Clark & Choi (2005) an experimental approach was adopted to compare the two interfaces on measures of interface performance and learning outcomes. Based on the persona effect, we expected that the chat-based interface would be associated with higher evaluation ratings and better learning outcomes in comparison to the text-based interface.

Study Description

Piagetbot was built using most of the same design decisions and content associated with Freudbot. Students using this interface could 'chat' with Jean Piaget by typing natural language questions and comments in a text box and having the agent respond with an appropriate answer. Like Freudbot, Piagetbot was designed to output autobiographical or conceptual narratives in 'chunks' as each reponse would contain implicatives that direct the user to continue the conversation. Paigetbot also had conditional strategies to direct the user to specific topics when input could not be recognized. Finally, there was also a photo image of Jean Piaget displayed on the side and partially animated to appear as though the lips were moving as new output was presented. There was also smoke occasionally wafting from a depicted pipe.

The text-based interface provided content about Piaget organized by chronological events and major theoretical constructs. Students using this interface could navigate the site using the site map presented on the side. It is important to note that the same content was used in both interfaces with the exception of first vs. third person references to Piaget. As well, the navigational menu that accompanied the text-based interface was also displayed in the chat-based interface so that participants in both conditions were given the same overview of topics available to review or discuss.

Students were told that the purpose of the study was to compare two types of web interfaces on a student's ability to learn information about Jean Piaget and his theory of cognitive development in children. After providing informed consent, participants completed a brief 7-item pretest to assess their existing knowledge of Jean Piaget and his developmental theory. After the pretest, they were randomly assigned to one of the two interfaces and given 20 minutes to learn all they could about Piaget in preparation for a 35-item learning outcome test that would immediately follow the review. After the test, participants completed a short questionnaire where they were asked to evaluate their interface experience and provide information on relevant individual difference variables. After the questionnaire was completed, participants were provided with an electronic debriefing and thanked for their participation.

There were 59 participants in the present study (9 men and 50 women). The participants were Athabasca University students currently or recently registered in one or more psychology courses. Forty-four percent participants listed themselves as full-time students and 52% as part-time. Participants were recruited into the study primarily through an email request from the first author. A monetary incentive to participate was also provided (1/30 chance in a draw for $300).

Results

Table 1 displays the average ratings for each of the questions used to evaluate the interface experience on a 5-point scale where higher values were associated with positive experiences. To examine whether elements of the interface experience were rated differently between the two groups, t-tests were used to compare ratings. Levine's test of homogeneity of variance was also calculated and the degrees of freedom were adjusted in cases where the assumption of homogeneity was violated. The ratings for the two conditions are shown in the right half of Table 1. Interestingly, the group of participants in the chat-based interface rated their experience significantly lower across 5 of the 7 questions than students in the text-based condition. There were no differences between interfaces on ease of use and degree of engagement

Interface Evaluation	Text Interface n=28	Chat Interface n=31

	M	SD	M	SD	t
How enjoyable was this activity?	3.43	0.69	2.87	1.12	2.32*
How engaging was this activity	3.43	1.03	3.19	1.22	0.79
How easy was this activity?	3.11	0.96	3.19	1.11	0.32
Would you recommend this activity to others?	3.68	0.77	2.97	1.33	2.54*
Overall, how would you rate this activity?	3.43	0.74	2.87	1.11	2.23*
How useful is this interface for learning information about Jean Piaget?	4.22	0.84	3.42	1.31	2.80**
How useful is this interface for remembering information about Jean Piaget?	3.53	0.74	3.00	.89	2.49*

Table 1: Interface evaluation ratings based on a 5-point scale where higher numbers reflect positive experiences. Note: * - $p < .05$; ** - $p < .01$

An analysis of scores from the pretest of Piagetian knowledge indicated that although participants in the text-based interface had higher scores (M = 46.9%, SD=19.8%) than did participants in the chat-based condition (M = 38.3%, SD=20.7%), the difference was not statistically significant, t (57) = 1.65, ns. Before analyzing data from the learning outcome test, an inspection of the digital record indicated that the average participant in the text-based interface was exposed to about 85% of the web content in contrast to average participant in the chat-based interface who only saw 25% of the web content. To adjust for differences in exposure, the learning outcome measure was calculated using only questions that were associated with relevant exposed content. Not surprisingly, the percentage of eligible questions (out of 35) was higher for the participants in the text-based interface (M = 83.79%, SD=21.1%) than did participants in the chat-based condition (M = 56.3%, SD=19.8%), a difference that was statistical significant, t (57) = 5.14, p < .001. The learning outcome measure, defined as the percent of correctly answered eligible questions, was analyzed in an Analysis of Covariance with Condition as a fixed factor and the pretest score as a covariate. There was no effect of the covariate but there was a significant effect of Condition, F (1, 56) = 5.59, p < .05. Participants in the text-based condition had significantly higher learning outcome scores (M = 67.4%, SD=13.9%) than did participants in the chat-based condition (M = 56.8%, SD=16.2%). The same analysis was carried out on the percent of correctly answered ineligible questions and as one might expect, performance was considerably poorer. Importantly, the difference between the text-based condition (M = 26.8%, SD=28.49%) and the chat-based condition (M = 33.0%, SD=19.9%) was non significant, F (1, 56) = 1.75, ns.

In order to evaluate the performance of Piagetbot, data from the same questions used in Heller et al. (2005) to evaluate Freudbot were compared to the data associated with Piagetbot. As shown in Table 2, both Freudbot and Piagetbot were equally rated and despite the differences in sample sizes, there were no statistical differences between the agents on any of the 5 evaluation measures. Moreover, although there were slightly fewer participants that indicated they would chat with Piagetbot again compared to Freudbot (58% vs. 68%), there was no statistical difference in the proportion of repeat chatters for each agent, χ^2 (1, N = 84) = .828, ns.

Agent Evaluation	Freudbot n=53		Piagetbot n=31		
	M	SD	M	SD	t
How enjoyable was this activity?	2.92	1.05	2.87	1.12	0.22
How engaging was this activity	3.08	1.11	3.19	1.22	0.45
Would you recommend this activity to others?	2.83	1.12	2.97	1.33	0.52
Overall, how would you rate this activity?	3.02	0.93	2.87	1.11	0.65
How useful is this interface for learning information about Sigmund Freud/Jean Piaget?	2.96	0.94	3.42	1.31	1.70

Table 2: Ratings based on a 5-point scale where higher numbers reflect positive experiences. Note: All t's were non-significant

Discussion

The results indicated that under well controlled experimental conditions, the text-based interface was evaluated more favourably and was associated with better learning outcomes than was the chat-based interface even after the outcome measure was adjusted to account for differences in content exposure. These findings were initially unexpected but comments on the outgoing survey alerted us to an important limitation of the study in relation to the orienting instructions given at the outset. Students were told that a multiple choice test would follow each interface. Because of past education experiences involving exam preparation, the text-based interface was likely the best format to fit the instructions Exam preparation by chatting may have been too novel or in this case, too inefficient for effective performance. In retrospect, the text-based interface clearly provided the most efficient access to content. This observation was supported by the comments of many students who expressed frustration with the inability of Piagetbot to provide all of the information all at once. One possible remedy to this limitation is to use task orienting instructions that are compatible with the strengths of each interface coupled with an incidental test of memory as opposed to intentional recall.

Another important limitation in the present study is the lack of animation in our conversational agent leading to weak test of the engagement mechanism hypothesized to underlie learning outcome effects. Dirkin, Mishra & Altermatt (2005) recently argued that the persona effect may only occur when the agent is fully social and animated (or interestingly, when the agent is absent). As reported earlier, both Clark & Choi (2005) and Dehn & van Mulken (2000) acknowledge that variability in the animation of pedagogical agents may be partly responsible for inconclusive findings. Gulz and Haake (2006) also argue that more attention needs to be devoted to the look of pedagogical agents. In our current work, the conversational agents are being integrated with interactive character software to create full animation historical figure applications.

Finally, the findings indicated a good transfer of historical application design principles as evaluations of Piagetbot were indistinguishable from Freudbot and suggest a wider application base within and across disciplines may be feasible. Altogether, the present findings indicate that the role of conversational agents in distance education needs to be carefully examined with attention paid to pedagogical goals and outcomes.

References

André, E., Rist, T, & Müller, J. (1999). Employing AI methods to control the behaviour of animated interface agents. *Applied Artificial Intelligence,* 13, 415-448.

Atkinson, R.K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94(2), 416-427.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchhill (Eds.), *Embodied Conversational Agents* (pp. 29 – 63). Cambridge: MIT Press

Clark, R.E., & Choi, S. (2005). Five design principles for experiments on the effects of animated pedagogical agents. *Journal of Educational Computing Research,* 32(3), 209-225.

Craig, S.D., Driscoll, D.M., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia*, 13(2), 163-183.

Dehn, D.M., & van Mulken, S. (2000). The impact of animated interface agents: A review of the empirical literature. *International Journal of Human-Computer Studies,* 52, 1-22.

Dirkin, K.H., Mishra, P., & Altermatt, E. (2005). All or nothing: Levels of sociability of a pedagogical software agent and its impact on student perceptions and learning. *Journal of Educational Multimedia and Hypermedia*, 14(2), 113-128.

Fosnot, C. (1996). Constructivism: A Psychological theory of learning. In C. Fosnot (Ed.) *Constructivism: Theory, perspectives, and practice*, (pp.8-33). New York: Teachers College Press.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R. & Tutoring Research Group (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1, 35-51.

Gulz , A., & Haake, M. (2006). Design of Animated pedagogical agents: A look at their look. *International Journal of Human-Computer Studies,* 64(4), 322-339.

Hadwin, F.H., Winne, P.H., & Nesbit, J.C. (2005). Roles for software technologies in advancing research and theory in educational psychology. *British Journal of Educational Psychology*, 75, 1-24.

Heller, R.B., Procter, M., Mah, D., Jewell, L., & Cheung, B. (2005). Freudbot: An Investigation of Chatbot Technology in Distance Education. *Proceedings of the World Conference on Multimedia, Hypermedia*, and Telecommunications.

Heo, H. (2004). Story telling and retelling as narrative inquiry in cyber learning environments. In C. M. D. J.-D. &. R. P. R. Atkinson (Ed.), *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference*.

Johnson. W.L., Rickel, J.W., & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education,* 11, 47-78.

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., & Bhogal, R.S. (1997a). The persona effect: Affective impact of animated pedagogical agents. In S. Pemberton (Ed.), *Human Factors in Computing Systems:CHI'97 Conference Proceedings.* New York: ACM Press.

Mayer, R.E., Dow, G.T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds. *Journal of Educational Psychology*, 95(4), 806-813.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents. *Cognition and Instruction,* 19(2), 177-213.

Moundridou, M., & Virvou, M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, 18, 253-261.

Rumelhart, D.E (1975). *Notes on schema for stories.* In Daniel G. Bobrow and Allan Collins, editors, Representation and Understanding: Studies on Cognitive Science, Language, Thought, and Culture: Advances in the Study of Cognition, pages 211-236. Academic Press, New York.

Thorndyke, P. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology,* 9, 77-110.

Acknowledgements

This research was supported by a grant from Athabasca University Mission Critical Research Fund to the first author.