The Impact of Conversational Navigational Guides on the Learning, Use, and Perceptions of Users of a Web Site

Art Graesser[1], G. Tanner Jackson[1], Matthew Ventura[1], James Mueller[2], Xiangen Hu[1], Natalie Person[2]

[1] University of Memphis, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230,
{a-graesser, gtjacksn, mventura, xhu} @memphis.edu
[2] Rhodes College, Department of Psychology, Memphis, TN, 38112
{muejn, person} @rhodes.edu

Abstract. Knowledge management systems will presumably benefit from intelligent interfaces, including those with animated conversational agents. One of the functions of an animated conversational agent is to serve as a navigational guide that nudges the user how to use the interface in a productive way. This is a different function from delivering the content of the material. We conducted a study on college students who used a web facility in one of four navigational guide conditions: Full Guide (speech and face), Voice Guide, Print Guide, and No Guide. The web site was the Human Use Regulatory Affairs Advisor (HURAA), a web-based facility that provides help and training on research ethics, based on documents and regulations in United States Federal agencies. The college students used HURAA to complete a number of learning modules and document retrieval tasks. There was no significant facilitation of any of the guides on several measures of learning and performance, compared with the No Guide condition. This result suggests that the potential benefits of conversational guides are not ubiquitous, but they may save time and increase learning under specific conditions that are yet to be isolated.

1 Introduction

Knowledge management systems are expected to be facilitated by intelligent interfaces that guide users who vary in cognitive abilities, domain, knowledge, and computer literacy. Some users will not have the patience to learn systems that are not used very often. These users will need fast and easy guidance. Some prefer to talk with agents in a conversational style rather than reading dense printed material on a computer screen and typing information via keyboard. Therefore, there has been serious interest in intelligent interfaces that have speech recognition and animated conversational agents. These agents incorporate synthesized speech, facial expressions, and gestures in a coordinated fashion that attempts to simulate a conversation partner. An ideal interface would allow the user to have a conversation with the computer, just as one would have a conversation with a person.

Animated conversational agents have been explored in the context of learning environments and help systems during the last decade [2], [3], [4], [11], [12], [14],

L. van Elst, V. Dignum, and A. Abecker (Eds.): AMKM 2003, LNAI 2926, pp. 48-56, 2003.

[19]. There is some evidence that AutoTutor, a tutoring system with an animated conversational agent, improves learning when college students learn about computer literacy or conceptual physics by holding a conversation with the computer tutor [11], [21]. However, it is still unsettled what aspects of a conversational agent might be effective, and under what conditions [2], [19], [23]. Is it the voice, the facial expressions, the responsiveness to the user, the gestures, the content of the messages, or some combination of these features? Whittaker (2003) has concluded that the voice is particularly effective in promoting learning and in engaging the user's attention, but the other components of the agent may be effective under specific conditions that are not yet completely understood.

One potential function of an animated conversational agent is to serve as a navigational guide to offer suggestions on how the user might use the interface in a productive way. This is an entirely different function from delivering the content of the material that would otherwise be read. The purpose of the present study was to investigate different types of conversational navigational guides that are available to adults when they use a new web site. Do these guides saving time for the user when the agents offer suggestions on what to do next? Does the user acquire more information because of the time that is allegedly saved? What are the perceptions of users toward conversational navigational guides? Do the like them, or are the suggestions irritating? It is widely acknowledged that the Microsoft's Paperclip irritated many users because of its intrusiveness and the difficulty of getting rid of it. Perhaps a better designed, more conversationally appropriate, agent would be more appreciated by the user.

We conducted a study on 155 college students who used a web facility in one of four navigational guide conditions: Full Guide (speech and face), Voice Guide, Print Guide, and No Guide. The web site was the Human Use Regulatory Affairs Advisor (HURAA), a web-based facility that provides help and training on the ethical use of human subjects in research, based on documents and regulations in United States Federal agencies [9]. The college students used HURAA to complete a number of training modules and document retrieval tasks.

2 Different Types of Navigational Guides

The Full Guide was a talking head with synthesized speech, facial expressions, and pointing gestures. The Agent told the user what to do next when the user first encountered a web page. For example, when the user entered the "Explore Issues" module, the Agent said, "Select the issue that you would like to explore." The talking head also moved to direct the user's attention to some point on the display. For example, the talking head looked down when he said "You may select one of the options below me." The talking head told the user what each primary and secondary module was supposed to do, after the user rested the mouse pointer over a module link for more than 2 seconds. The Agent was designed to project an authoritative persona and to help the user navigate through the interface more quickly. Many novice users are lost and don't know what to do next when they encounter a page. The Agent was designed to reduce this wasted time.

In order to directly test the influence of the Agent as a navigational guide, participants were randomly assigned to one of the following four conditions:

Full Guide. There is the full talking head.
Voice Guide. There is a voice that speaks, but no head.
Print Guide. The guidance messages are printed at the location where the talking head normally is.
No Guide. There are no messages of navigational guidance, either spoken or in print.

If a navigational guide is important, then the completion of the various tasks should be poorer in the No Guide condition than the other three conditions: $d < min\{a, b, c\}$. When considering the three conditions with the guidance, there is a question of what medium is effective. If speech reigns supreme, then $c < min\{a,b\}$. This would be predicted by available research that has compared the impact of spoken versus printed text on comprehension and memory [2], [19], [23]. If print is superior, then the prediction would be that $c > max\{a,b\}$. If the presence of the face provides a persona effect that improves interactivity [14], then the prediction is $a > b$. However, if the face is a distraction from the material in the main display, then the prediction is $a < b$.

3 Human Use Regulatory Affairs Advisor (HURAA)

HURAA is a web-based facility that provides help, training, and information retrieval on the ethical use of human subjects in research. The content of HURAA is derived from Federal agency documents and regulations, particularly the National Institutes of Health [20], the Department of Defense [6], [7], and particular branches of the US military. The targeted users of HURAA focus on fundamental ethical issues, but not the detailed procedures and paper work associated with gaining approval from Institutional Review Boards.

The design of HURAA was guided by a number of broader objectives. The layout and design of the web facility incorporate available guidelines in human factors, human-computer interaction, and cognitive science [5], [17]. The architecture of the HURAA components needed to be conformant with the ADL standards for reusable instructional objects, as specified in the Sharable Content Objects Reference Model [22]. The primary objective of having these standards is to allow course content to be shared among different lesson planners, computer platforms, and institutions. HURAA was designed to optimize both learning and information transmission. Adult users are likely to have very little time, so it is important to optimize the speed and quality of learning in web-based distance learning environments. This requires careful consideration of the pacing of the information delivery, the selection of content, and design of the tasks to be performed. The web site was supposed to be engaging to the use, so there was persuasive multimedia intended to hook the user to continue on the website. Finally, HURAA incorporated some of the sophisticated pedagogical techniques that have been implemented in advanced learning environments with intelligent tutoring systems and animated conversational agents.

HURAA has a number of standard features of conventional web facilities and computer-based training, such as hypertext, multimedia, help modules, glossaries, archives, links to other sites, and page-turning didactic instruction. HURAA also has more intelligent features that allegedly promote deeper mastery of the material, such as lessons with case-based and explanation-based reasoning, document retrieval though natural language queries, animated conversational agents, and context-sensitive Frequently Asked Questions (called *Point & Query*, [10]). Additional details about HURAA can be found in Graesser, Hu et al., 2002. This paper directly focuses on some of the tasks users would complete with HURAA and what impact the 4 different guides had on the completion of these tasks and the users' perceptions of the learning environment.

4 Materials and Procedure

The experiment included three benchmark tasks that participants completed while interacting with HURAA. This was followed by a series of tests and surveys that were completed after they interacted with HURAA. We refer to these two phases as the HURAA **acquisition phase** and the post-HURAA **test phase**, respectively. The next section describes the modules and HURAA facilities that are directly relevant to the performance evaluation. The participants were 155 undergraduate students at the University of Memphis and Rhodes College who participated for course credit or for money ($20).

4.1 HURAA Acquisition Phase

Introduction. The Introduction Module is a multimedia movie that plays immediately after a new user has logged in. It is available for replay for users who want to see a repeat. The Introduction is intended to impress the user with the importance of protecting human subjects in research. It introduces the user to the basic concepts of the Common Rule [6], [20], of the Belmont Report's coverage of beneficence, justice, and respect for persons, and of the Seven Critical Issues that must be scrutinized when evaluating any case [8]: Social and scientific value, accepted scientific principles, fair subject selection, informed consent, minimizing risks and maximizing benefits, independent review, and respect for subjects. The Introduction was prepared by an accomplished expert in radio and web-based entertainment industries, after rounds of feedback from a panel of DoD personnel.

Lessons. This module has four lessons that teach the user about the Seven Critical Issues identified by Emmanuel et al. (2000) and how to apply them to particular cases that involve ethical abuses. This is a form of case-based reasoning [1], [15]. The first lesson presented the user with descriptions of the Seven Critical Issues, a summary of the Tuskegee Syphilis Study, and an explanation of how each of the Seven Critical Issues was violated in the Tuskegee study. The second lesson presented the user with a description of a study on post traumatic stress disorders. The user was then presented with the Seven Critical Issues and must decide, on a six-point scale, the extent to which there potentially is a problem with each issue in that case. The six point scale is: 1 = Definitely not a problem, 2 = most likely not a problem, 3 = undecided, guess it's not a problem, 4 = undecided, guess it's a problem, 5 = most likely a problem, and 6 = definitely a problem. The user then received feedback comparing his/her responses with those of a panel of experts from the DoD, along with a brief explanation. Discrepancies between the learner's decisions and the judgments of the experts were highlighted. Lesson 3 followed the same procedure as Lesson 2, except there was another case on a routine flight test with an experimental helmet. Lesson 4 presented two additional cases, following the same procedure. One was on helmet-mounted devices and the other on chemotherapy.

Signal detection analyses were performed on the learner's decisions as a measure of performance. There are four categories of decisions when signal detection analyses are applied.

Hit (H). Both the learner and expert agree that an issue is potentially problematic for the particular case.

Correct rejection (CR). Both the learner and expert are in agreement that an issue is **not** potentially problematic for a case.

Miss (M). The expert believes there is a potential problem, but the learner does not.

False alarm (FA). The learner believes there is a problem, but the expert believes there is no problem.

The experts were 7 experts on research ethics in the military. A ***d'* score** was also computed that assesses how well the learner can discriminate whether a case does versus do not have a problem with respect to an issue. A ***d'* score** of 0 means the learner is not at all discriminating whereas the score increases to the extent that the user is progressively more discriminating (with an upper bound of about 4.0).

Query Documents. This module allows the user to ask a natural language question (or description) and then generates an answer by retrieving high matching excerpts from various documents in the HURAA web site. For each document that the user selects, the highest matching paragraph from the document space is selected by the computational linguistics software and is displayed in a window. Beneath this window, the headings for the next four results appear. If the top choice is not the one that the user needs, s/he can click on the headings to read those excerpts. The search engine that was available to identify the optimal matches was latent semantic analysis [16], [13].

In the **search task**, the participants were instructed to search the document space in order to find answers to 4 test questions. The participants recorded the answers in a test booklet. If the answer to a question was lengthy, they were instructed to write down the fetched document and section number where the answer was found. Performance was measured by retrieval time and the likelihood of retrieving the correct paragraph out of the large document space. If the natural language query facilities are useful, then there should be facilitation in the speed and likelihood of accessing the correct documents.

4.2 Tests in the Test Phase

The test consisted of three parts: (1) **Memory** (a test on the important ideas from the Introduction and Lessons), (2) **Issue comprehension** (a test on the participant's ability to identify potentially problematic issues in cases), and (3) **Perception ratings** (ratings on how the participants viewed the learning experiences).

Memory for Important Ideas. This phase tests memory for the central, core ideas from the Introduction and Lesson material. These core concepts are those that all users should take away from the learning experience. Memory was assessed in three subtests: Free recall, cued recall, and the cloze task. The free recall test presented a series of concepts that the participants were asked to define or describe off of the top of their head. After finishing the free recall task, the cued recall test was administered on the next page. The cued recall test had more retrieval cues than the free recall test.

The cloze procedure has the most retrieval cues. It took verbatim segments of the introductory text and left out key words, which the participant filled in. There were progressively more retrieval cues for content to be retrieved as one goes from free recall to the cloze task.

Issue Comprehension. This test assessed how discriminating the participants were in identifying potentially problematic issues on two cases. The cases were selected systematically so that 6 of the issues were problematic in one and only one of the two cases; one of the issues was problematic in both cases so it was not scored. This test is functionally a transfer test from the case-based, explanation-based reasoning task in the HURAA acquisition phase. The participants simply read each case and rated the seven issues on the 6-point scale (as to whether issue I was problematic for case C.

Perception Ratings. The participants gave ratings on their perceptions of the learning environments. The four rating scales that were included in all three experiments are presented below. The values on each rating scale were: 1 = disagree, 2 = somewhat disagree, 3 = slightly disagree, 4 = slightly agree, 5 = somewhat agree, and 6 = agree. Examples are as follows: You learned a lot about human subjects protections." and "It was easy to use and learn from these instructional materials."

5 Results and Discussion

Table 1, on the last page, presents means and standard deviations of the dependent measures in the four experimental conditions. The most striking finding from the experiment is the lack of significant differences among conditions. In fact, there were no significant differences among the conditions for any of the 13 dependent measures in Table 1. This null result is incompatible with all of the above predictions. It should be emphasized that the sample size was quite large so the likelihood of a type II error was not high.

The practical implication of the result is that the animated conversational agent did not facilitate learning, usage, and perceptions of the interface. In essence, the agent and the conversational guidance had no bang for the buck. Perhaps the web facility was designed extremely well, so well that a navigational guide was superfluous. The navigational agent might prove to be more effect when the information on the screen is more complex, congested, and potentially confusing. Knowledge management systems often have complex information displays so the value of these agents may increase as a function of the complexity, ambiguity, and perplexity of the system. Perhaps there are special conditions when a navigational guide of some form will be helpful, whether it be print, voice, or a talking head. However, these precise conditions have yet to be discovered and precisely specified in the literature.

It is appropriate to acknowledge that the results of the present study on agents as navigational guides does not generalize to other learning environments. Animated conversational agents have proven to be effective when they deliver information and learning material in monologues and tutorial dialogues [2], [19], [21], particularly when the test taps deep levels of comprehension. However, only a handful of empirical studies has systematically investigated the impact of these conversational agents on learning, so more research is definitely needed. One intriguing finding is that the amount of information that a person learns and remembers from a learning system is not significantly correlated with how much the learner likes the system [18]. Simply put, learning is unrelated to liking. It this result is accurate, then it is not sufficient to simply ask users and individuals in focus groups what they like or do not like about agents and navigational guides. There also needs to be a serious, deep, research arm that goes beyond intuitions of users, designers, and managers.

6 Acknowledgements

This research was directly supported by contracts from the Office of Naval Research (N61339-01-C1006) and the Institute for Defense Analyses (AK-2-1801), and was partially supported from grants by the National Science Foundation (REC 0106965) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR, IDA or NSF.

References

1. Ashley, K.D. (1990). Modeling legal argument: Reasoning with cases and hypotheticals. Cambridge, MA: MIT Press.
2. Atkinson, R.K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416-427.
3. Baylor, A.L. (2001). Investigating multiple pedagogical perspectives through MIMIC (Multiple Intelligent Mentors Instructing Collaboratively). *Workshop on agents at the AI-ED International Conference*. San Antonio, TX.
4. Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. AI Magazine, 22, 67-83.
5. Collins, A., Neville, P., & Bielaczyc, K. (2000). The role of different media in designing learning environments. International Journal of Artificial Intelligence in Education, 11, 144-162.
6. 32 CFR 219 (1991). Protection of Human Subjects, Department of Defense.

7. DoD Directive 3216.2 (1993). Protection of Human Subjects in DoD supported research, Department of Defense.
8. Emmanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? Journal of the American Medical Association, 283*,* 2701-2711.
9. Graesser, A.C., Hu, X., Person, N.K., Stewart, C., Toth, J., Jackson, G.T., Susarla, S., Ventura, M. (2002). Learning about the ethical treatment of human subjects in experiments on a web facility with a conversational agent and ITS components. In S. A. Cerri (Ed.), Proceedings of Intelligent Tutoring Systems 2002 (pp. 972-981). Berlin, Germany: Springer.
10. Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), The psychology of learning and motivation: Vol. 29. Categorization by humans and machines (pp. 411-436). Orlando, FL: Academic Press.
11. Graesser, A.C., Person, N., Harter, D., & TRG (2001). Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education, 12, 257-279.
12. Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. AI Magazine, 22, 39-51.
13. Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8, 129-148.
14. 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.
15. Kolodner, J. (1984). Retrieval and organizational strategies in conceptual memory: A computer model. Hillsdale, NJ: Erlbaum..
16. Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.
17. Mayer, R.E. (1997). Multimedia learning: Are we asking the right questions? Educational Psychologist, 32, 1-19.
18. Moreno, K.N., Klettke, B., Nibbaragandla, K., Graesser, A.C., & the Tutoring Research Group (2002). Perceived characteristics and pedagogical efficacy of animated conversational agents. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems 2002* (pp. 963-971). Berlin, Germany: Springer.
19. 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 the interact with animated pedagogical agents? *Cognition & Instruction, 19*, 177-213.
20. NIH 45 CFR 46 (1991). National Institutes of Health Code of Federal Regulations, Protection of Human Subjects.
21. Person, N.K., Graesser, A.C., Bautista, L., Mathews, E., & TRG (2001). Evaluating student learning gains in two versions of AutoTutor. In J.D. Moore, C.L. Redfield, and W.L. Johnson (Eds.) Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future (pp. 286-293). Amsterdam: OIS Press.
22. Sharable Content Object Reference Model (SCORM), Versions 1.1 and 1.2. www.adlnet.org.
23. Whittaker, S. (2003). Theories and methods in mediated communication. In A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (Eds.). Handbook of discourse processes. Mahwah, NJ: Erlbaum.

Table 1. Means (and Standard Deviations) for Dependent Measures.

DEPENDENT MEASURES	Full Guide	Voice Guide	Print Guide	No Guide
Number of participants	40	39	38	38
Memory for Core Concepts				
Free recall proportion	.45 (.21)	.43 (.20)	.42 (.21)	.44 (.20)
Cued Recall proportion	.51 (.24)	.50 (.23)	.45 (.26)	.53(.23)
Cloze recall proportion	.44 (.15)	.42 (.18)	.39 (.17)	.47 (.19)
Introduction study time (minutes)	6.6 (3.3)	9.3 (19.0)	7.4 (4.7)	10.7 (19.8)
Problematic Issue Identification				
Hit proportion	.58 (.10)	.58 (.10)	.56 (.12)	.60(.13)
False alarm proportion	.40 (.32)	.39 (.30)	.38 (.34)	.38(.31)
d' score (discrimination)	.30 (.46)	.31 (.36)	.14 (.75)	.27 (.74)
Task completion time (minutes)	23.7 (6.2)	23.7 (8.1)	20.8(5.2)	22.9(4.1)
Search for Information				
Correct document retrieval (Proportion)	.55(.25)	.50 (.49)	.47(.23)	.52(.25)
Search time (minutes)	27.1(11.4)	24.4(9.6)	23.5(9.2)	24.8(7.9)
Perception ratings				
Amount learned	4.75(1.06)	4.62(1.16)	4.61(1.08)	4.53(1.48)
Interest	3.85(1.63)	4.08(1.46)	4.11(1.41)	3.79(1.82)
Enjoyment	3.50(1.47)	3.46(1.48)	3.58(1.41)	2.89(1.47)
Ease of learning	4.13(1.40)	3.95(1.45)	3.71(1.71)	3.61(1.57)