An Empirical Study of the Effect of Agent Competence on User Performance and Perception

Technical Report GIT-GVU-04-03
February 2004

Jun Xiao[1], **John Stasko**[1], *and* **Richard Catrambone** [2]
[1]College of Computing / [2]School of Psychology & GVU Center
Georgia Institute of Technology
Atlanta, GA 30332 USA
{junxiao,stasko}@cc.gatech.edu, rc7@prism.gatech.edu

Abstract

We studied the role of the competence of a user interface agent/assistant that helped users to learn and use a new text editor. Participants in the study made a set of prescribed changes to a document via the editor with the aid of one of four interface agents. Participants could ask questions out loud to the agent and the agent would respond using a synthesized voice; the agent would also make proactive suggestions. The agents varied in the quality of responses and suggestions made. One group of participants were provided with a help screen as well as the agent. We focused on assessing the relation between users' objective performance, interaction style, and subjective experience. Results revealed that the perceived utility of the agent was influenced by the types of errors made by the agent, while participants' subjective impressions of the agent related to the perceptions of its representation. In addition, allowing participants to choose their preferred assistance style(s) (agent vs. online-help) improved objective performance. We correlate quantitative findings with qualitative interview data and discuss implications for the design and the implementation of systems with interface agents.

Keywords

Embodied conversational agents, interface agent, empirical evaluation, qualitative study, quantitative study

1 Introduction

The creation of interface agents that present relevant information, promote active learning, and cooperate with users to perform tasks is a complex endeavor that requires research in a variety of areas [13]. An enormous amount of effort has been spent on building sophisticated interface agent systems in many application domains such as entertainment, health care, and education [1,2,9]. Much research of this type has been presented at various conferences and workshops [5,11,15,16,19]. On the one hand, we have perceived continuous improvement in the quality of the planning and reasoning of disembodied agents who work intelligently behind the scene. On the other hand, we have witnessed increasing believability of the personality and affect of embodied conversational agents who interact with users with the whole bandwidth of different modalities.

While the potential of interface agents demonstrated in the laboratory is compelling, attempts to realize this potential in practice so far, such as Microsoft's Clippy in the Office software, have failed to deliver on the promise, a point acknowledged by Microsoft itself [14]. The challenge to find the delicate balance between providing proactive help and interrupting users in an annoying way remains a significant obstacle to the wide-scale adoption of interface agents. Therefore, we suggest that, rather than exploit a particular technology, researchers shift the focus to a more holistic perspective on the evaluation of such systems and remain grounded in real situations. Being a relatively immature area, the development of interface agent systems truly needs methodical support for the iterative process of gathering requirements, measuring results and refining design [10]. The purpose of our work, therefore, is to contribute to the HCI research community's understanding of interface agents by conducting empirical studies that identify the dimensions of the design space, uncover the correlations and tradeoffs between factors impacting agent design, and highlight areas for improvement.

This paper calls attention to one fundamental issue with the quality of interface agents, competence[1]. Developers of interface agent systems often dismiss this issue as an implementation issue and examine it by simply taking traditional metrics and assessing overall system performance. The investigation of the impact of the agent's competence on users presents a challenge of its own because of the complex interplay of various components of an interface agent system. Furthermore, because many of these studies tend to be informal, the results often lack internal and external validity.

We present a study that examined the effect of an agent's competence on user's objective performance and subjective experience via both concrete quantifiable measures and in-depth, qualitative session analysis. The rest of the paper is

[1] We use the term "competence" to refer to the general quality, appropriateness, and timeliness of help provided by an interface agent. The perception of competence is clearly subjective, but here we use the term in a more objective manner. If a user asks an agent what is "2+2", a competent answer is 4. Similarly, if an agent suggests a course of action that is relevant and appropriate to a user's situation, we view that suggestion as being competent.

organized as follows. In the next section, we briefly discuss results from prior studies on interface agents that provide unsolicited help. Then we introduce the primary research goals of this empirical study, followed by the description of the experimental setting. Next, we give a detailed analysis of results and summarize findings with respect to design concerns and implementation issues of interface agents. Finally, we conclude this paper with directions for future study.

2 Related Work

Rich and his colleagues introduced a collaborative model underlying the COLLAGEN system for interface agents to provide both tutoring and assistance [18]. The initiative of the agents was adjustable with different parameter settings. However, what parameter values to choose and how to change those values over time remained an open question.

Horvitz adopted a decision theoretic approach to regulate the behavior of interactive agents designed to provide unsolicited help [7]. He suggested a set of heuristics be followed to balance the utility of the information with the cost of disruption it would cause the users. However, none of the heuristics was empirically validated, but rather based on intuition.

Mackay conducted an experiment comparing two groups of participants who received relevant suggestions or random ones [12]. She found that participants learned the same number of commands for a text editor, regardless of the relevance of the suggestions to the participants' immediate activities. However, because she used a "yoked control", the participants in the yoked condition who supposedly only got irrelevant information might also have received relevant information.

Randall and Pedersen examined various approaches applied by several help systems to assist users with problem solving [17]. They argued that mediated help provided by interface agents was either enjoyable and helpful or cutesy and annoying depending on whether the users could recognize and understand the ability of the agents. However, the basis of the argument was personal experience and thoughts rather than scientific evidence from publicized empirical studies.

3 Experiment

In a prior study, we investigated the use of interface agents to assist users in an editing task [20]. The agents provided appropriate responses to user questions and made timely proactive suggestions, and thus were competent. Participants using the agents performed as well as those using a printed manual, which was arguably the best help one could have for such limited, straightforward tasks, and felt that the agent was useful.

In this study, we examine how degraded levels of agent competence and different styles of assistance affect user performance and impression. The rationale for examining degraded performance is that current agents are far from competent and therefore it is important to examine user reactions to less-than-competent agents. The quality of the replies and suggestions made by the interface agents is manipulated through a Wizard of Oz simulation [4]. Data about participants' objective performance, interaction style, and subjective experience were collected and analyzed.

While interface agents certainly have more potential than just providing procedural help for a text editing program, we believe this focused study could provide us a wealth of opportunities for exploring the following practical questions and the results could be extended to a broad class of interface agent systems:

- How will degradation of interface agents' competence affect user performance and perceived usefulness of the agents?
- Will the competence of an interface agent and the performance of a user affect the likeability of the agent?
- Will user preferences, especially preference of assistance style, have an effect on user performance and subjective assessment? Given alternative choices (online-help), will people use an agent at all?

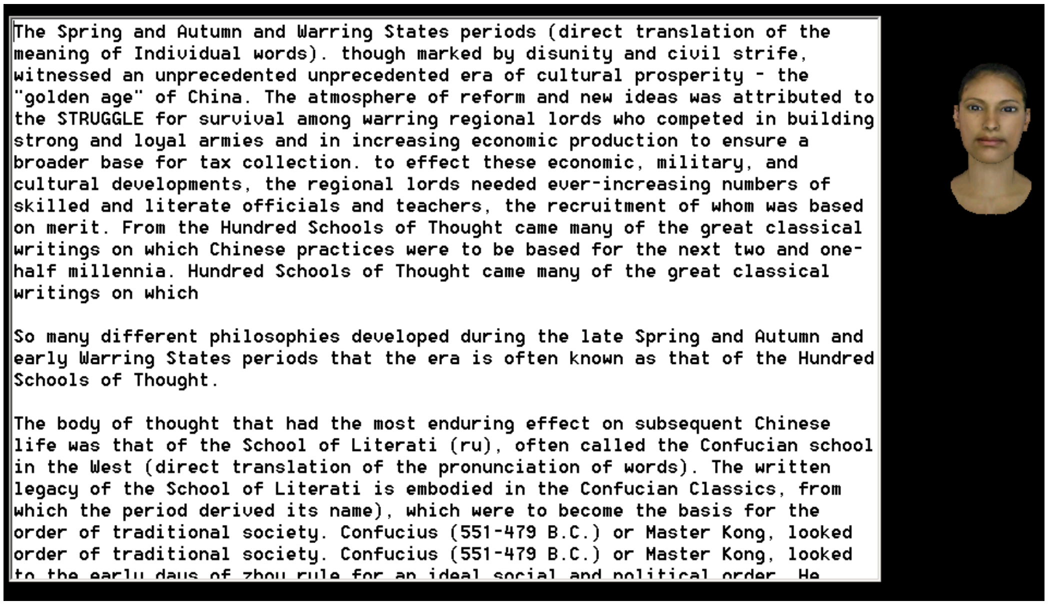
3.1 Experimental Setting

Participants

Fifty-one Georgia Tech undergraduates (31 male, 20 female) with a variety of majors and computer experience were randomly assigned to conditions and received course credit for participating. Analysis of demographic information collected during the experiment showed that gender was approximately balanced across conditions and none of these participants was familiar with command-language based editors similar to the one described below.

Software and Equipment

We built a text editor for the experiment. In order to keep the interface simple and clear, no menus or other interface objects were available besides the text window, the



The Spring and Autumn and Warring States periods (direct translation of the meaning of Individual words). though marked by disunity and civil strife, witnessed an unprecedented unprecedented era of cultural prosperity - the "golden age" of China. The atmosphere of reform and new ideas was attributed to the STRUGGLE for survival among warring regional lords who competed in building strong and loyal armies and in increasing economic production to ensure a broader base for tax collection. to effect these economic, military, and cultural developments, the regional lords needed ever-increasing numbers of skilled and literate officials and teachers, the recruitment of whom was based on merit. From the Hundred Schools of Thought came many of the great classical writings on which Chinese practices were to be based for the next two and one-half millennia. Hundred Schools of Thought came many of the great classical writings on which

So many different philosophies developed during the late Spring and Autumn and early Warring States periods that the era is often known as that of the Hundred Schools of Thought.

The body of thought that had the most enduring effect on subsequent Chinese life was that of the School of Literati (ru), often called the Confucian school in the West (direct translation of the pronunciation of words). The written legacy of the School of Literati is embodied in the Confucian Classics, from which the period derived its name), which were to become the basis for the order of traditional society. Confucius (551-479 B.C.) or Master Kong, looked order of traditional society. Confucius (551-479 B.C.) or Master Kong, looked to the early days of zhou rule for an ideal social and political order. He

Figure 1. Interface that participants encountered.

agent, and the black background (see Figure 1). The mouse was disabled during the experiment session. To issue an editing command, a participant had to press a specific control or escape key combination. Text entry occurred for any other key selections. This restricted command language style design enabled us to better interpret and predict participants' actions and proactively give appropriate help.

The editor offered a rich set of commands for high-level operations on words, lines, and sentences in addition to basic character-oriented commands. To accomplish the editing task, the participant likely would to need to consult the agent or online manual for help. However, it was not our intention to over-burden the participant with too many keystrokes to remember. Thus, pilot testing was conducted to ensure that while the number of commands (30) was sufficiently large, the keystrokes for the commands were not too demanding to learn and recall.

An experimenter was present to introduce the participant to the experiment materials and to guide the participant in using the computer equipped with a microphone and speaker. Another experimenter (the wizard) in an adjacent room listened to the questions asked by a participant and monitored the participant's computer screen and editing activities via composite video images from two cameras. The entire session was recorded.

We developed a software monitoring tool to assist the wizard with the simulation of the decision making of the interface agent. The editing tasks performed by participants were decomposed into elementary sub-tasks and appropriate advice was associated with each task. The wizard caused the agent to produce responses using the monitoring tool. In addition, every keystroke issued by the participants and interactions between the participant and the interface agent were automatically logged.

Training

At the beginning of the experiment session, a participant watched a short video tutorial that showed the editor interface and explained the usage of each commands as a sample document was being edited. Every keystroke combination needed to issue the various commands was described in the following order: cursor movements, text modification, and font style change. Participants were advised to watch the video closely but they were informed that they would be able to get help from a computer agent (and could also use the online help screen in the *Online-Help* condition) when carrying out editing tasks. After showing the video, the experimenter initiated the editing program and demonstrated how to use the computer agent by asking the agent a sample question: "How do I insert text into a document?" (and explained how to use the help screen in the *Online-Help* condition).

Editing Tasks

Each participant was instructed to make the same set of 22 prescribed changes to an existing document using the text editor. Pilot testing was conducted to ensure that the tasks were of appropriate difficulty and that participants using a help sheet could usually complete the editing task in less than 30 minutes. Participants were required to perform the tasks in order and encouraged to revise the document as accurately and as quickly as they could.

The sequence of editing tasks was composed carefully so that certain more powerful short-cut commands (e.g., delete a sentence) would likely be used twice with other commands mixed in between. This was done so that if a participant asked the interface agent or consulted the online help at the first opportunity to use a short-cut command, the second occurrence would later allow the participant to apply such knowledge. In addition, if proactive advice was given to remind a participant of the existence of short-cut commands, it would increase the chance that the participant would use the command (perhaps rather than some less efficient command, if one existed, such as deleting a sentence character by character) later at the second occurrence of such an opportunity.

3.2 Interface Agent

An earlier study of ours suggested that user performance was not affected by different appearances (e.g., realistic, stiff, or iconic) of interface agents when conducting simple objective tasks, such as an editing task, but people felt more comfortable talking to an interface agent with a life-like appearance [3]. Therefore, the interface agent (from Haptek [6]) we chose to use in this experiment had a realistic animated 3D female appearance; the agent blinked and moved its head occasionally (see Figure 1) in addition to moving its mouth in synchronization with synthesized voice from the Microsoft Speech SDK. The female appearance was chosen from a set of possibilities because of more neutral responses from several people during pilot tests.

The agent stayed in a small window, uncontrollable by the participant, at the upper-right corner of the computer screen. No intonation, speech style, gaze patterns, or facial expressions were applied by the agent to convey emotion or personality. The agent always answered questions and gave advice in the following predefined manner: “To X press Y” (e.g., “To delete a character, press control-d”). We will discuss the effect of those design decisions on the participants later in this paper.

As mentioned earlier, the agent’s responses were controlled through a Wizard of Oz technique. The wizard had full knowledge of all the editing commands and gave the best answer for the participant’s question based on the keywords spotted in the question. A variety of responses covering other situations were also prepared. These included responses such as asking the participant to repeat the question (such as when background noise was louder than the participant’s voice) or to rephrase the question (such as when there were two or more matches for the spotted keywords) or to state that the agent was not able to provide an answer (such as when a participant asked for a function that the editor did not possess).

In addition to answering questions, the agent sometimes offered unsolicited help; participants were informed this might occur. In order to make the proactive advice from the wizard resemble that of a plausible interface agent with respect to performance and consistency, the interface agent was not omniscient. Rather, proactive advice was offered only when it was logically plausible for a computer agent to make inferences from the participant's actions in the following two situations:

First, if a participant performed a subtask inefficiently, for example, moving the cursor character-by-character to the beginning of the sentence instead of applying the move-to-the-beginning-of-a-sentence command, then the agent would suggest the key sequence for the short-cut command if the intention of the participant was clear. Second, if the agent could logically predict what the next task would be, for example, a copy would likely be followed by paste operation, the agent would tell the participant the key combination needed for that very next task (e.g., it would tell the participant the key sequence for the paste command right after the participant had copied some text). The agent did not explain the reasoning for its suggestions.

3.3 Conditions

The experiment was a five-group, between subjects design. The first four conditions varied the competence of the agent that responded to questions and made proactive suggestions. Participants in the fifth condition were able to use the online-help as well as the agent.

Competent Condition

The interface agent in this condition always answered a command query with the appropriate, correct key sequence for the command. It also offered suggestions only at relevant times. Furthermore, for the purpose of being less disruptive to a participant, the agent would suggest a command only if the participant had not previously used or asked about it; a particular suggestion would be made at most once.

Moderate Competence Reactive Condition

Similar to the agent in the Competent condition, the agent in this condition offered proactive suggestions only at relevant times. However, the agent made mistakes when answering the participant's questions. For every five questions, the agent gave two inappropriate replies with the constraint that it would give the correct answer to immediately repeated questions.

We found in pilot sessions that if the incorrect answer did not relate to the question (e.g. reply with how to move forward a word command if asked about how to delete a line), it raised the participant's suspicion that the responses were intentionally manipulated rather than generated. Therefore, we constructed the inappropriate replies so that they would appear to be reasonable errors. For example, if the participant asked how to under*line* a word, the inappropriate reply for this question would be how to move down a *line*.

In order to reduce variability of the agent's performance, the same inappropriate reply was provided for the same question across all participants in this condition.

Also, all participants received incorrect answers in the same randomly generated sequence. Finally, we carefully inspected the prepared inappropriate reply in the context of editing tasks to avoid unintentionally giving away useful information to the participant. For example, if a participant asked how to move down a line and needed to move to the end of document for the very next task, the wrong answer for the question did not say how to move to the end of the document.

Low Competence Reactive Condition

Similar to the agent in the Moderate Competence Reactive condition, the agent in this condition made relevant proactive suggestions, but gave inappropriate replies. However, the agent gave inappropriate replies to participants' questions at a higher rate than in the Moderate Competence Reactive condition. The agent gave inappropriate replies half of the time, again in a predetermined randomly generated sequence. Unlike the Moderate Competence Reactive condition, the agent could repeat incorrect answers to consecutive repeated questions as well.

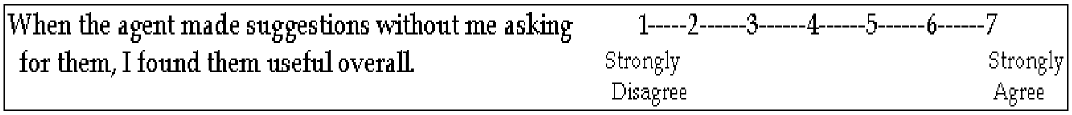
Low Competence Proactive Condition

In contrast to the agent in the Low Competence Reactive condition, the agent in this condition answered all the participant's questions appropriately, but made irrelevant proactive suggestions half the time (e.g., when the participant was deleting a line by characters, instead of suggesting the delete-a-line command, the agent suggested how to move down a line).

Again, for reasons of consistency, all participants in this condition received irrelevant proactive suggestions in the same random sequence. We inspected the prepared irrelevant suggestions in the context of editing tasks to avoid accidentally providing helpful information. For example, if a participant had never asked how to move to the end of the document and needed to move to the end of document for some upcoming task, the irrelevant suggestion was not about how to move to the end of document at that moment.

Online-Help Condition

In addition to using the Competent condition's interface agent, participants in this condition could use online-help. By toggling the F1 function key, the participants could switch between the help screen and the editor screen. All 30 commands were listed on one screen in two columns. To find the key sequence for a particular command, the participant just browsed through the list of commands in four directions using arrow keys to move to the desired command. The appropriate key sequence then would appear to the right of the command. This design allowed us to track which command the participant was looking for but still keep the help screen easy and familiar.



When the agent made suggestions without me asking for them, I found them useful overall. 1-----2------3------4-----5------6------7 Strongly Disagree Strongly Agree

Figure 2. Example of Questionnaire Item

3.4 Assessment

Satisfaction Questionnaire and Free Response Interview

After completing the editing tasks, participants filled out a seven-point agree-disagree Likert format questionnaire (see an example statement in Figure 2) about their subjective experience with the computer agent and the editor. In addition, the experimenter asked a few open-ended questions about participants' impressions of the agent and related issues at the end of each session.

Performance Measures

Using the software log, we measured the amount of time participants spent on the editing tasks and how many keystrokes, except those involving text entry, the participants pressed to carry out the tasks. We also tallied how many times participants received help from the agent, reactively and proactively (and from the help screen in the *Online-Help* condition).

4 Results and Discussion

Below, we outline the answers to our earlier three research questions by describing findings from the quantitative data supplemented with qualitative interview data. We conclude each section with discussion of the implications for the design and implementation of interface agents.

4.1 Perceived Utility of an Agent Varies with Types of Errors the Agent Made

We asked the participants, using the Likert survey, whether they found the agent's responses to questions to be useful. An ANOVA was carried out to determine whether the reactive competence of the agent we manipulated had an effect on the ratings. It turned out that there was a significant difference among conditions (competent: 6.4, moderate: 5.8, low reactive: 4.6, low proactive: 6.2, online-help: 6.3); $F(4, 46) = 5.32$, $MSE = 1.21$, $p = .001$. The agent's responses in the Low Competence Reactive condition were rated as significantly less useful than in the Competent, Low Competence Proactive, and Online-Help conditions (all *p*s < .02), and marginally less useful than in Moderate Competence Reactive condition ($p = .064$). There was no difference between the Competent and Moderate Competence Reactive conditions ($p = .37$).

One participant talked about how frustrated she was with such an agent during the interview: "It (the agent) was not helpful. The only thing was that it had a hard time recognizing some voice commands. It had really poor vocabulary, just couldn't understand what I was saying." All but one participant in the Low Competence

Reactive condition voiced similar complaints about the usefulness of the agent. It seems clear that this agent, with a response error rate of 50%, was found to be less useful.

On the other hand, we found a very different perception with respect to proactive suggestions. We asked participants whether they found the unprompted suggestions to be useful. Analysis of the ratings for this item showed no significant effect of condition (competent: 4.6, moderate: 5.3, low reactive: 5.1, low proactive: 4.2, online-help: 5.0); *F*(4, 46) = 0.65, *MSE* = 3.29, *p* = .63. Furthermore, during the interview we asked the participants in the Low Competence Proactive condition what they thought about the fact that the agent sometimes made suggestions without them asking for help. None of the participants in the Low Competence Proactive condition gave a definite "No, it's not helpful" answer. Rather, typical comments were, "Most of time it was helpful," "I found it to be helpful sometimes," and "It was all right. Sometimes it was useful, sometimes it was not so useful." When further asked about what they thought about the irrelevant suggestions, participants explained, "The usefulness far exceeded the occasional annoyance," and "Each time I didn't need that information, but it may be helpful." One participant put it this way: "It was trying get inside my head and trying to understand what I was doing. Whether it is a human being or computer, it would never be 100% correct when you try to do that. Maybe I was fumbling with the keys [and] gave the computer the wrong impression of what I was trying to do."

Although the agent in Low Competence Proactive condition had a similar 50% error rate in its suggestions just as the agent in the Low Competence Reactive condition had with its answers, participants' perceptions of the usefulness of the suggestions varied far less. Our findings indicated that errors in reactive answers and proactive suggestions weigh differently on people's perceptions of the usefulness of an interface agent. People seem to be more forgiving with proactive errors than reactive errors because they do not expect proactive suggestions to always be "right."

Participants in the Moderate Competence Reactive condition, when asked how they felt about getting incorrect answers, often surprisingly felt that they were not receiving incorrect answers: "[I didn't] phrase my questions right," "I wasn't speaking loudly and clearly enough," and "The word I was saying could be confusing," One participant described why errors occurred: "It was probably looking for keywords. If I didn't convey my question the way it would be able to understand, it would do its best to give me whatever answer it had based on the information it had. So I just went ahead and rephrased my questions. It did respond correctly when I asked the right questions."

It appeared that as long as the agent did not make "repeated errors", as one participant in the Low Competence Reactive condition referred, people would blame themselves rather than the agent. One participant put it this way: "When I phrased the question in what seems like the best way to phrase it to me, she (agent) knew the answer. There wasn't much confusion of me having to say the questions over and over." Another participant even characterized the situation as the agent being able to correct itself: "I asked pretty much the same question twice, it gave me different answers because it figured out that the first answer it gave me wasn't what I was looking for, so maybe gave me related answers as well." When asked about further suggestions for improving the agent, one participant said, "not necessarily the agent, maybe just preparing the person, telling them what type of commands she understands." In short, using the words of one participant, those in the Moderate Competence Reactive condition seemed to feel this way about the agent: "She was definitely helpful, she was just not perfect."

Returning to considerations of proactive suggestions, further findings indicate more of a homogeneous view of proactive advice. Nine of the 31 participants in the Competent, Moderate Competence Reactive, and Low Competence Reactive conditions (all of whom used an agent that gave only appropriate proactive advice according to our definition), perceived some of the agent's suggestions as being irrelevant. In the final Likert survey, we asked the participants to characterize whether they thought the proactive suggestions came at the right time. There was no significant difference between the Competent and Low Competence Proactive conditions for this item (competent: 4.9, low proactive: 4.2); $p = 0.36$. A common reason we found during the interview (4 of 9 participants made this comment) was that participants thought the advice should have come earlier. One participant argued, "Normally, it tells you to fix something, which is a good idea, but if it's two seconds earlier, it would be useful. I thought it made the suggestion too late most of the time after I'd already done it the long way." Another issue some participants (three of the nine) had with the proactive suggestions was that the participants needed to focus on the current task and the suggestions were a form of distraction. One participant made the argument that the advice was irrelevant because "it wasn't what I needed at that time. I found it useful if I knew it was coming. But if I was in the task already, in the middle of doing something, it was kind of distracting." Finally, two participants took supposedly good proactive advice as being irrelevant because they actually did not realize or notice the suggestions were relevant. One participant explained, "At first I thought it was annoying, but later a couple of times, I realized that it was actually useful." In short, some participants thought the proactive suggestions were not so helpful even if they were seemingly timely and relevant.

Although the results of this study can be generalized only with caution to other application domains, the findings discussed above suggest some important implications for the development of applications with interface agents.

- Avoid repeating errors to greatly improve people's perceptions of the quality of reactive agents.
- People's expectations and perceptions of the usefulness of proactive agents are relatively low.
- Proactive suggestions are more readily accepted if they can be immediately applied and are easy to understand.

4.2 A Person's Subjective View of an Agent has Little to Do with its Utility

Our second research question concerns the relation between the utility of an agent and its subjective appeal. An ANOVA showed that none of the ratings for whether the agent was intrusive, friendly, intelligent, or annoying (means for the four ratings: 2.73, 4.51, 4.53, 2.71, with 1 = strongly disagree and 7 = strongly agree) was significantly different across the conditions (all *p*s > .10). Furthermore, tests on the correlation for the above ratings and participants' objective task performance (completion time and total number of keystrokes) were non-significant (see Table 1). Therefore, it appears that the usefulness of the agent had little effect on subjective views of the agent.

	Time	Keystrokes	Face	Voice
Intrusive	.04	.09	.28*	.17
Friendly	.16	.05	.47*	.33*
Intelligent	-.12	-.25	.37*	.24
Annoying	.05	.10	-.01	-.26*

Note: *N*=51. *p < .05

Table 1. Correlation for Satisfaction Ratings and Performance and the Perceived Quality of Face and Voice

Actually, given that the perceived usefulness of proactive suggestions differed little across conditions, it is not surprising to observe that participants did not find the agent in the Low Competence Proactive condition to be more annoying than the one in the Competent condition. In fact, most participants had positive reactions towards the proactive suggestions: "It was kind of nice that they (developers) openly did that", "They (suggestions) weird me out because they were exactly what I am thinking", and "That (suggestions) was what I was most impressed with". Four participants who were not in the Low Competence Proactive condition and thus received competent suggestions did notably disagree, however. They simply described the proactive behavior of the agent as "interfering" and "annoying". One participant put it this way: "I don't think I was listening. I heard what it said, but I just didn't use those commands (suggestions). The whole concept is that you are controlling everything. Have that (agent) come in, even more time it is helpful than not, it will cause you like, 'ok, I don't want to mess with it,' even if it would help you in the long term."

Curiously, however, we found that participants' reactions to the agent related more to the perceived quality of the face and the voice of the agent. During the interview we asked the participants what they thought about the face and the voice of the agent and asked them to rate the quality. It turned out that whether participants viewed the agent as friendly, intelligent or intrusive was all positively correlated with the quality of the face (see Table 1). Similarly, the quality of the voice was positively correlated with participants' perceptions of the agent being friendly and negatively correlated with perceptions of the agent being annoying.

In fact, it is quite striking to see the divergent effects of the same face between participants. Roughly half of the participants had favorable reactions while the other half had negative comments. One participant said, "It's strange to look at it", whereas another participant observed, "The face is pretty friendly looking". One participant described, “I don’t like the way she looked, a stoned person, looked like a zombie, somebody on drugs”, whereas another participant argued, “She didn’t need to move constantly. She can just move whenever you talk to her”. Finally, while one participant claimed, “I tried to ignore (the face)”, another participant stated, “I might glance every now and then when I didn’t know what was going on.”

Similarly, participants’ impressions of the voice varied greatly. One participant described, “The voice with it was really computer-ish, dull, sounds more like a machine with its monotone voice”, whereas another participant expressed, “It’s pleasant and helpful from the tone”. Another participant with a negative view stated, “(The voice was) haunting. (It would) keep me up at night. The sound just annoyed me because it’s monotone and unnatural.” But a more positive participant said, “It was not annoying. The voice was fine. It sounded like a machine, but it didn’t drone like a machine or (something) purely electronic”.

During the interview, as a concluding question, we asked participants their impressions of the Microsoft Office paper clip assistant “Clippy”. The vast majority of responses were negative. One participant described the only moment when he was happy with it: “The thing is annoying as hell and hard to get rid of. It always comes back! At one point, though, I typed in ‘how do I make you go away?’ and it did told me how to make it go away. That is pretty sweet.” Five participants made comparisons with the agents in our study. One participant thought Clippy didn’t look real: “Because the paper clip doesn’t look like a real person, you don’t really feel like asking for help from it.” Four participants expressed dissatisfaction with Clippy because it uses a dialog that must be attended to as opposed to the agent in our study that could easily be ignored. One participant put it this way: “It (Clippy) gets in the way because it uses text instead of voice. It shows up, might be blocking some of the tools. You have to keep dragging it around. Using voice, it (the agent) doesn’t block the tool and I can do what the agent tells me to do or just ignore it without having to divert from the task.” As a final note, four participants said they liked the MS Office assistant after they changed its appearance. One participant said, “I changed it to a cat. It’s pretty much just there to amuse me, not to help. It’s kind of, my entertainment.” Another participant simply put it this way: “I changed it to the dog. It’s cute. I like my dog.” Another participant simply put it this way: “I changed it to the dog. It’s cute. I like my dog.”

The findings discussed above should not be taken as a blanket discount of the importance of improving the competence of the agent, but rather it calls attention to several factors when designing applications that use an embodied interface agent.

- The subjective appeal of a proactive interface agent has more to do with people’s individual differences than with the competence of the agent.
- People’s subjective experience with an interface agent relates to features of the agent, such as face and voice.
- It is unlikely that one can design a representation of an interface agent that would suit all users.

4.3 Preferred Assistance Styles Relate to Performance

As mentioned before, we measured the number of keystrokes pressed by a participant as an indicator of how efficiently the participant was performing the editing tasks. Our analysis showed that there was a significant difference across conditions (*F*(4, 46) = 2.73, *MSE* = 128587.60, *p* = .04) on this measure. More specifically, participants in the Online-Help condition significantly out-performed other participants in terms of efficiency (all pairwise *p*s < .05). None of the other pairwise comparisons showed a significant difference (*p*s > .2). In addition, participants in the Online-Help condition tended to perform better in terms of completion time (see Table 2), but the variability across individuals in each group eliminated a significant effect.

	Commands	Time
Competent	972 (430)	1446 (418)
Low Competence Reactive	867 (279)	1407 (232)
Moderate Competence Reactive	955 (505)	1612 (606)
Low Competence Proactive	843 (338)	1414 (326)
Online-help	519 (103)	1245 (199)

Table 2. Means (Standard-Deviations) for Number of Commands Issued and Completion Time (in sec.)

Further inspection of data from the 10 participants in the Online-Help condition revealed that three exclusively used the agent, three exclusively used the help screen, and four used both the agent and the help screen roughly equally. Our prior study showed that the participants with competent agents performed as well as the participants who had a manual instead of an agent and even performed a little better in terms of efficiency [20]. Similarly, performance of the participants in this study's Online-Help condition varied little as a function of whether they used the agent or the help screen. However, the fact that participants in the Online-Help condition could *choose* and use their preferred assistance style, appeared to be the reason for the overall performance improvement.

During the interview, we asked these participants why they preferred the agent or the help screen or used them both. Of the three participants who used the help screen exclusively, one participant characterized the help screen as easy and direct to access: "All the information was on one screen and it was easy to get to. It's always there. Whenever I need it, I can see it and it tells me what exactly I need to do." Another participant thought that using the help screen was faster than the asking the agent: "I thought it would be more efficient just to, not have to worry about speaking to the computer and waiting for the response when I could just do it in two seconds." The third participant just did not feel comfortable asking the agent for help: "I prefer the help screen. I feel pretty dumb, in most situations, talking to a computer."

Interestingly, the participants who exclusively used the agent made similar arguments the other way. One participant said the agent was easy to use: "I tried to use it (the help screen) the first time. It took too long so I just asked the questions the rest of time. It was convenient, you can just say it out loud and get an answer back and I trust the agent." Another participant noted that it was more efficient to use the agent: "It gave me the ability to edit and ask questions and get answers at the same time. I don't have to switch between screens in terms of leaving the document, going to the help screen, looking at what I need and then going back." The third participant just preferred asking questions: "It was easy to hit a key and use the help screen, but saying is a more natural act, at least for me."

The participants who used both the agent and the help screen talked about their mixed reactions. One participant used the agent after she found out that the agent was actually helpful and said, "I didn't want to talk to her (the agent) at first. Once she started talking, it's like ok. I used it quite a bit." Later, when asked about her opinion about the Microsoft office assistant, she said, "I only tried it once, and it didn't seem that useful. I end up not using it. I won't ever use it any more. This one (the agent) was much better at figuring out what I was asking. She was better at telling me. The one on my computer (Clippy) just jumps in and tries to sell itself, like 'I notice you are writing a letter. Can I help?'" It appears that prior impressions played a key role in this person's behavior.

Three other participants also took advantage of both the agent and the online-help. One participant said, "Sometimes as I used it I remembered seeing that specific help topic. It's easy to hit [F1] and look and check really fast. She (the agent) would be useful, but I don't know if it's worth the time." Another participant explained, "I used the agent when I wanted to navigate the document. I wouldn't ask the agent so much for commands of hot keys for formatting because I can reference those easily on the help screen." The other participant simply said, "I would use the agent for more complex questions that would be more time consuming to reference on the help screen."

As a final note, we would like to mention that one participant in the Competent condition who performed the worst in that group described his experience this way: "I am definitely a very visual person. Whenever the instructions were given at the beginning (tutorial), I was kind of like 'What was going on?' I can't comprehend all those different directions. If it was written down, I can see it and look at it, it's right there. It would be more productive if I were able to memorize the commands from a manual than having to ask the agent every occasion. And I don't know how to phrase it in computer terms or not."

Although this study examined two particular interaction styles with a specific application, the findings discussed above imply some potentially important design guidelines for applications using interface agents. Further investigation of people's preference with respect to interaction styles is clearly needed.

- It is crucial to match a user's preferred interaction style to the way help is provided.
- Provide alternative forms of help rather than just an agent acting as an assistant.

- It is important to build people's confidence by illustrating the utility of interface agents.

5 CONCLUSIONS AND FUTURE WORK

This paper provides an empirical test that is particularly important under the cloud cast on the future of interface agents by the Microsoft Office assistant. The results showed that the apparent failure of Clippy should not be taken as a blanket rejection of the use of interface agents. To the contrary, echoing Isbister and Nass [8], we believe continuing efforts to address problems of matching user characteristics with features of agents will yield significant enhancements in human-agent interaction. Future research should extend the current study with additional manipulations of proactive suggestions, crossing all agent conditions with online-help, and having participants work on more complex tasks in longitudinal studies.

6 Acknowledgements

Haptek Inc. has generously provided software and technical support for this research.

7 References

[1] Andre, E., Rist, T., van Mulken, S., Klesen, M, & Baldes, S. The automated design of believable dialogues for animated presentation teams. In *Embodied Conversational Agents*, Cassell, J., Sullivan, J., Prevost, S. & Churchill, E. (eds.), 2000, MIT Press, 346-373.

[2] Bickmore, T. Relational agents: Effecting change through human-computer relationships. Ph.D. Thesis, 2003, MIT.

[3] Catrambone, R., Stasko J., & Xiao, J. Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research. Proc. of CogSci 2002, 166-171.

[4] Dahlback, N., Jonsson, A., & Ahrenberg, L. Wizard of Oz studies – Why and how. Proc. of IUI 1993, ACM Press, 193-200.

[5] Diederiks, E. Buddies in a box: animated characters in consumer electronics. Proc. of IUI 2003, ACM Press, 34-38.

[6] Haptek. URL: http://www.haptek.com.

[7] Horvitz, E. Principles of mixed-initiative user interfaces. Proc. of CHI 1999, ACM Press, 159-166.

[8] Isbister, K. & Nass, C. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. International Journal of Human-Computer Studies, 53(1) (2000), 251-267.

[9] Johnson, W. L, Rickel, J. W., & Lester, J. C. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education, 41 (2000), 41-78

[10] Johnson, W. L., Shaw, E., Marshal, A., & LaBore, C. Evolution of user interaction: the case of agent adele. Proc. of IUI2003, ACM Press, 93-100.

[11] Lieberman, H. Autonomous interface agents. Proc. of CHI 1997, ACM Press, 67-74.

[12] Mackay, W.E. Does tutoring really have to be intelligent? Ext. Abstracts CHI 2001, ACM Press, 265-266.

[13] Maes, P. Agents that reduce work and information overload. Communications of the ACM 37, 3 (1994), 31-40.

[14] Microsoft PressPass. URL: http://www.microsoft.com/presspass/features/2001/apr01/04-11clippy.asp.

[15] Pelachaud C., Carofiglio V., De Carolis B., De Rosis, F., & Poggi, I. Embodied contextual agent in information delivering application. Proc. of AAMAS2002, ACM Press, 758-765.

[16] Pelachaud, C., Ruttkay, Z., Marriott, A. AAMAS Workshop Embodied Conversational Characters as Individuals. 2003.

[17] Randall, N., Pedersen, I. Who exactly is trying to help us? The ethos of help systems in popular computer applications, Proc. of the 16th annual international conference on Computer documentation 1998, ACM Press 63-69.

[18] Rich, C., Lesh, N.B., Rickel, J., Garland, A. A Plug-in Architecture for Generating Collaborative Agent Responses. Proc. of AAMAS2002, ACM Press, 782-789.

[19] Rist, T., Aylett, R., Ballin, D., Rickel, J. (eds.). 4th International workshop on intelligent virtual agent. Proc. of IVA 2003. Springer.

[20] Xiao, J., Catrambone, R., and Stasko, J. Be quiet? Evaluating proactive and reactive user interface assistants, Proc. of INTERACT 2003, IOS Press, 383-390.