# The Role of Choice and Customization on Users' Interaction with Embodied Conversational Agents: Effects on Perception and Performance

**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 performed an empirical study exploring people's interactions with an embodied conversational agent (ECA) while performing two tasks. Conditions varied with respect to 1) whether participants were allowed to choose an agent and its characteristics and 2) the putative quality or appropriateness of the agent for the tasks. For both tasks, selection combined with the illusion of further customization significantly improved participants' overall subjective impressions of the ECAs while putative quality had little or no effect. Additionally, performance data revealed that the ECA's motivation and persuasion effects were significantly enhanced when participants chose agents to use. We found that user expectations about and perceptions of the interaction between themselves and an ECA depended very much on the individual's preconceived notions and preferences of various ECA characteristics and might deviate greatly from the models that ECA designers intend to portray.

**Author Keywords**
Embodied conversational agents, interface assistants, empirical evaluation, qualitative analysis, controlled experiment, personalization, and customization.

**ACM Classification Keywords**
H.5.2. Information interface and presentation: User Interfaces – Evaluation/methodology.

**INTRODUCTION**
An Embodied Conversation Agent (ECA) is a computer-generated interactive character with human-like appearance and lifelike behavior that answers questions and performs tasks for the user through conversational, natural language-style dialogs [3]. The momentum behind ECA research comes from the desire to migrate from the "computer as a crowded tool-box" metaphor to the human guide metaphor.

This paper explicitly calls attention to one issue: how the option to personalize or customize an ECA affects both subjective impression as well as objective performance when a person works on tasks in conjunction with an ECA.

Earlier studies in this area indicate that individual differences play a major role in the acceptance or rejection of certain ECA features [1, 8, 9]. Thus, a "one size fits all" approach in deploying ECAs simply might not provide the flexibility required for allowing users to opt for their best match and to facilitate optimum utility and usability of the interface. In addition, human-ECA interactions utilize multiple modalities that demand abilities from users that vary greatly from individual to individual. Individual differences are likely to be accentuated by the very fact that ECA interfaces are full of uncertainties, especially in interpreting each other's beliefs, desires, and intentions. Because users cannot predict or understand all of an ECA's actions, they might develop expectations about the ECA's capability based upon their preconceived notions, form perceptions of the ECA's upon their own likings, and act and react accordingly.

Studies have been done on the effect of individual differences of users on their perception of ECA characteristics. For example, Nass' study of the effect of similarity attraction between human and ECA suggests that individual differences plays an important role [8]. However, user customization or selection of ECA features has been rarely explored except in games. In practice, developers of ECA applications often face the dilemma of whether to spend valuable resources building the "ideal" ECA. Surprisingly, no one has yet demonstrated formally that allowing users to choose certain features of ECAs could significantly improve both subjective impressions of the ECAs and objective task performance.

## EXPERIMENT OVERVIEW

This study presents a first-of-its-kind examination of the effect of choice and customization on the usefulness of ECAs. We sought to obtain a better understanding of personal variation in the expectations and perceptions of ECA interfaces and whether user customization can increase the effectiveness of ECAs. However, there is a major obstacle in this type of research: the study results might be influenced by the level of customization and the variety of ECAs available in an experiment. If we get a null result on customization effects, we cannot judge if it occurred because too little customization was provided or because the ECAs in the experiment were all poorly designed.

In this study, we reduced the above-mentioned obstacle by holding the ECA variable constant across conditions and creating the *illusion* of customization instead. Additionally, we measured how strong customization affects people's perception and behavior by comparing its effect to other factors. Specifically, four main issues are addressed in this study:

1. Will selection and the illusion of customization, e.g. allowing users to choose their preferred ECA characteristics, improve overall subjective impressions of ECAs?
2. Will people's subjective views of ECAs then influence their assessment of the task or the quality of interaction?
3. Can changes in people's subjective impressions of ECAs positively affect their behavior and objective task performance using the ECAs?
4. How difficult is it for ECA designers to achieve the desired effects of selection and customization?

To address these questions, we conducted a 2×2 between subjects experiment in which participants were asked to complete two tasks with help from an ECA. The first variable was whether the ECA was assigned by the experimenter or chosen by the participants. The second variable was the "quality" or appropriateness of the ECA, in terms of both appearance and ability, for the task domain.

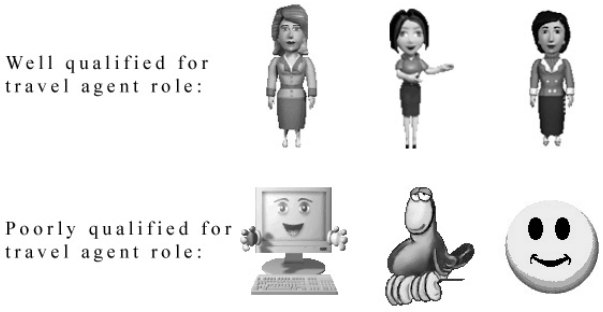


**Figure 1. Sample Microsoft Agent Characters.**

### Participants and ECAs

Sixty undergraduates (34 male, 26 female) with a variety of majors and computer experience selected from a general pool were randomly assigned to conditions and received course credit for participating. They were all native English speakers.

The ECAs used in the experiment were built with MS Agent SDK (see Figure 1). The ECAs were chosen from a pool of available characters based on a short survey (see Manipulation Check section below). Certain gaze patterns (e.g. glancing aside), facial expressions (e.g. smiling), and gestures (e.g. pointing to objects) were applied during the interaction when they were appropriate. The voice of the ECAs was generated using AT&T Natural Voices text-to-speech engine.

In order to factor out natural language understanding as an influence, the ECAs in the experiments were controlled through the Wizard of Oz technique. An experimenter was present to introduce the participants to the experiment materials and to guide the participants in using the computer equipped with a microphone and speaker. The participants were directed to speak clearly and loudly to the microphone when asking or answering questions with the ECAs. Another experimenter (the wizard) in an adjacent room listened to the questions and responses by participants and remotely controlled the scripted ECAs' responses and actions.

Interactions between the ECAs and the participants were all scripted. The wizard had no influence on the advice given by the ECA. The only variation introduced by the wizard was selecting one of several conversational conventions (e.g. "OK, let's continue") from canned responses for turn-taking purposes.

### Conditions

Participants were randomly assigned to one of the four conditions in a 2×2 balanced, between-subjects design: chosen ECA or assigned ECA × well qualified ECA or poorly qualified ECA. Participants performed two tasks about trip planning with the help of the ECAs.

With regard to the first independent variable, that is, chosen vs. assigned, in the two chosen conditions, the participants were allowed to choose an ECA out of a set of three possibilities. Each ECA would introduce itself and describe its experience. The participants were asked to choose which one they would like to interact with. In the two assigned conditions, the participants were simply assigned a predetermined ECA.

In addition, the participants in the two chosen conditions were presented with a further illusion of customization. At the beginning of the experiment, in the two chosen conditions, the participants filled out an online questionnaire about their preferences of ECA characteristics, for example, whether they would prefer a computer agent that spoke a little slow or fast or whether

they would prefer a computer agent that acts a little more formal or casual. The participants were instructed that their selection of the preferences would later determine which computer agents were available for the participants to choose from.

However, unbeknownst to the participants, the same set of three well qualified or poorly qualified ECAs were introduced to participants for selections. The ECAs were neutral in the dimensions of the participants' preferences so that we could create the illusion that the ECAs were customized to the preferences of the participants while at the same time achieving consistency across individual participants. Otherwise, the differences between the chosen ECAs would confound the study results.

With regard to the second variable, that is well qualified vs. poorly qualified, the ECAs in the well qualified conditions looked rather professional (see Figure 1), like a travel agent, and introduced themselves such as "I have been in the travel industry as an international travel consultant for more than ten years." In the poorly qualified conditions, the ECAs were represented as simple cartoons (see Figure 1), for example, a smiley face, and made self-introductions such as "I just got my first job as a receptionist at a travel agency last week."

### Pre-test Assessment

Given the literature on the relationship between user personality and preference for computer behavior [8], we suspected that the participants might respond and react differently based on their predispositions. Thus, we included composite measures for introversion and extroversion on the initial demographic questionnaire given to the participants based on Wiggins interpersonal adjective set. The index was very reliable (Cronbach's $\alpha = 0.91$).

Also, based on literature on the relationship between people's computer or technology anxiety and their performance with or perception of computer tasks [2], we suspected that participants might respond and react differentially based on their computer experience. Thus, we included a composite measure for computer experience on the demographic questionnaire based on the Computer Anxiety Rating Scale. The index was very reliable ($\alpha = 0.83$).

Finally, we were concerned that participants' perception of the usefulness of an ECA might also correlate with their domain knowledge (see below for task descriptions). Thus, we included two questions that measured the participant's travel domain knowledge on the demographic questionnaire: "Do you have travel abroad experience?" and "Will you be interested in a study abroad program?"

### Manipulation Check

Two different sets of ECAs were used for the two tasks. The well qualified and poorly qualified ECA characters were chosen based on an initial informal study. We showed all the ECAs to a set of people and asked them to rate the agents as being qualified or appropriate to serve as travel advisors. The top 6 out of 48 characters in the ratings were used in the experiment as well qualified ECAs for the two tasks and the bottom 6 were used in the experiment as poorly qualified ECAs. The descriptions of the ECAs were also independently examined by three raters to verify that the descriptions reflected the desired expertise of the ECAs, that is, well qualified vs. poorly qualified. With respect to participants' selections of the ECAs from the sets of three, no strong preferences were found and each ECA was selected by approximately an equal number of participants. This intentional manipulation ensures that the study results would not be confounded by the possibility that one particular ECA was more favorable than the others.

## SESSION ONE

### Trip Packing Task

In the first session of the experiment participants completed the International Trip Packing Problem. This problem involved a hypothetical situation in which the participant had a friend who was flying overseas on his first international trip. The task was to recommend six items for the person to take with him from a pool of 12 items and to rank the six items in order of importance.

The participants first did an initial ranking of the items. Then an ECA appeared and introduced itself to the participants in the assigned conditions. In the chosen condition, three ECAs introduced themselves to the participants and the participants were asked to pick one to work with. After that, the ECA made a predefined set of suggestions in which it recommended changing the rankings of four of the six items the participant selected and agreeing with the ranking of two other items. Finally, when the ECA finished providing all the feedback on the rankings, the participant did a final ranking of the items.

### Measures

#### *Subjective Measures*

Participant subjective impressions were measured by three sets of questionnaire items rated on 10-point Likert scales. Our choice of 10-point Likert was consistent with that done in other similar studies [8]. The first set of items assessed the participant's reactions to the ECA, for example, whether the ECA was attentive and competent. The second set of items examined the interaction between the participant and the ECA, for example, whether s/he enjoyed discussing the rankings of the trip packing items with the computer agent. The third set of items referenced the performance of the ECA and the participants themselves, for example, whether s/he did a better job on the ranking task with the computer agent than s/he would have without the computer agent.

Based on statistical factor analysis, four indices were created from the questionnaire items for the trip packing task: likeability, trust, usefulness and enjoyableness. All indices were highly reliable.

Likeability was an index created from four adjectives used to characterize the ECA: friendly, annoying, sociable, likable, and two statements/questions: "I would like to work with the computer agent in the future" and "How well did the discussion of the rankings with the computer agent go?" (Cronbach's $\alpha = 0.85$).

Trust was an index created from one adjective used to characterize the ECA: trustworthy, and two questions: "How much did you trust the advice from the computer assistant" and "How similar was your final rankings to the computer assistant's rankings suggestions?" ($\alpha = .82$).

Usefulness was an index created from one adjective used to characterize the ECA: helpful, three adjectives used to characterize the suggestions made by the ECA: constructive, worthwhile, thoughtful, and one statement: "I did a better job on the ranking task with the computer agent than I would have without the computer agent" ($\alpha = .88$).

Enjoyableness of the task was an index created from four adjectives used to characterize the task: boring, engaging, interesting, enjoyable, and one statement: "I enjoyed discussing the rankings of the items with the computer agent" ($\alpha = .87$).

One last question was used to measure participant's perceived task performance: "How well did you do the final ranking of the items?"

### *Objective Measures*

The likelihood of a participant following an ECA's advice is an interesting measure of the usefulness of an agent. While advice-following would certainly be at least partly a function of the quality of the advice, it was also affected by how the participants felt about the ECA. Therefore, ECA objective task performance, that is, persuasiveness, was also measured.

In this experiment, the ECAs were scripted to always comment that a participant's number one ranked item was not the most important item. Instead, it suggested promoting the participant's fourth ranked item to number one. Also the ECA would suggest removing the participant's fifth ranked item and selecting another item not on the participant's ranking list. In short, the ECA's suggestion was a predefined shuffle of the participant's ranking. The advice included both agreement and disagreement, and all participants received the same amount and level of disagreement and agreement.

After the discussion was over, participants were able to revisit the rankings and make changes. We measured whether participants changed their rankings as a function of the ECA's feedback by comparing the rankings of each item before and after the discussion between the participant and the ECA. Note that although the trip packing problem depends heavily on personal subjective preferences, the ECA's advice or suggestions were not based on the absolute value or importance of the listed items. Rather, the ECA's comments were based purely on the relative ranking order of the participants' choices.

| *Prefer an ECA that is* | *Percentage* |
|---|---|
| Quickly (vs. slowly) | 50% |
| Talkative (vs. terse) | 56% |
| Realistic (vs. cartoonish) | 46% |
| Diplomatic (vs. sarcastic) | 80% |
| Casual (vs. formal) | 83% |
| Extroverted (vs. introverted) | 83% |

**Table 1. Frequencies of Preferred ECA Characteristics by Participants in Chosen Conditions.**

Furthermore, we also conducted discourse analyses of participants' conversational behavior, for example, how many words per comment were made. We used this information to investigate potential behavioral and attitude differences across conditions.

## Results

### *Selections of ECA Characteristics*

Participants' selections of ECA characteristics in the initial selection process for the choice conditions are shown in Table 1. Consistent with the previous findings [9], participants had diverse opinions on the embodiment of ECAs. However, the data also showed that the majority of the participants preferred an ECA that spoke diplomatically over an ECA that spoke sarcastically, preferred an ECA that looked casual to an ECA that looked formal, and preferred an ECA that was extroverted to an ECA that was introverted. One possible explanation is that because the participants of this study were college students they would prefer to converse with people with similar characteristics, that is, casual. Also because in this experiment the ECAs acted as advisors, it is understandable that people would prefer an ECA that was diplomatic rather than sarcastic. Therefore, the nature of the task and the demographics of the participants seemed to imply that a diplomatic, extroverted and casual ECA would be viewed more favorable.

### *Subjective Impressions*

For all analyses, we performed full-factorial 2×2 ANCOVAs. Participants' Introversion-Extroversion Scale and Computer Anxiety Rating Scale were used as covariates.

We found significant main effects and interactions on the four indices for the trip packing task (see Figure 2).

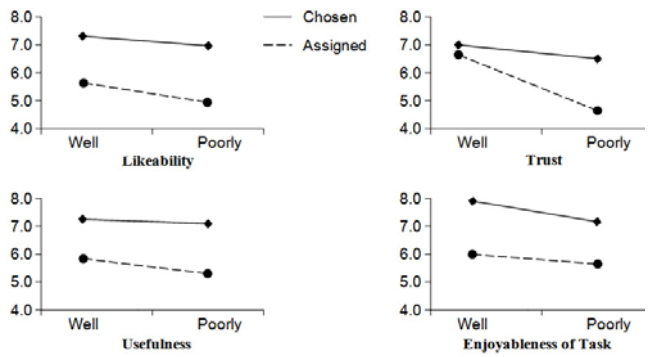


**Figure 2. Mean of Subjective Measures for the Trip Packing Task across Conditions.**

On likeability, there was a significant main effect of chosen vs. assigned; $F(1, 56) = 5.6$, $p < .02$. Participants in the chosen conditions viewed the ECA significantly more likable than participants in the assigned conditions.

On trust, there was a significant interaction between well qualified vs. poorly qualified and chosen vs. assigned; $F(1, 56) = 5.5$, $p < .02$. Posthoc pairwise Tukey tests identified that participants in the poorly qualified & assigned condition viewed the ECA as significantly less trustworthy than participants in the other conditions; all $p$'s < .05. No other pairwise comparisons reached significance level.

On usefulness, there was a significant main effect of chosen vs. assigned; $F(1, 56) = 4.9$, $p < .03$. Participants in the chosen conditions viewed the ECA significantly more useful than participants in the assigned conditions.

On enjoyableness of the task, there was a significant main effect of chosen vs. assigned; $F(1, 56) = 7.8$, $p < .01$. Participants in the chosen conditions viewed the ranking task significantly more enjoyable than did participants in the assigned conditions.

### *Task Performance*

As described earlier, participants were asked to rate how well they thought they did the ranking task. Additionally, the ECAs' task performance was measured by how closely the participants' final rankings were to the ECAs' suggestions. Finally, we counted how many words each participant used during the discussion with the ECA. Table 2 shows the results.

The first row of the table shows the participants' self-rated task performance across conditions. While there were trends to the differences across conditions, none of the pairwise comparisons reached significance at .05 level. As we suspected, the prior travel experience of the participants had a strong influence, however. We further did a stepwise multiple regression analysis, taking into account participants' self-rated travel abroad experience from the demographic questionnaire. The analysis showed that the chosen vs. assigned independent variable predicted the performance rating variable significantly over and above

| | Well Chosen | Well Assigned | Poorly Chosen | Poorly Assigned |
|---|---|---|---|---|
| Self-rated performance | 7.1 | 6.1 | 6.9 | 6.2 |
| Ranking agreement | 5.3 | 4.2 | 5.2 | 4.0 |
| Number of words | 481 | 423 | 458 | 437 |

**Table 2. Mean of Objective Measures of Trip Packing Task across Conditions.**

the participant's travel abroad experience rating variable; $R^2$ change = .17, $F(1, 56) = 7.1$, $p < .01$. Therefore, if participants had the same amount travel abroad experience, the participants in the chosen conditions would be more confident about their final rankings than participants in the assigned conditions.

The second row of the table shows the average number of items that the participant's final rankings agreed with the rankings suggested by the ECA across conditions. Participants in the chosen conditions agreed significantly more with the ECA ranking list than participants in the assigned conditions did; $F(1, 56) = 4.4$, $p < .04$ (this significance was reached by taking into account the influence of the participant's personalities and computer experience using an ANCOVA). Considering that the amount and level of disagreement and agreement between the ECA's ranking and the participant's initial rankings were the same across conditions, we can conclude that the ECAs in the chosen conditions were significantly more influential and persuasive than the ECAs in the assigned conditions. In contrast, the well qualified vs. poorly qualified variable manipulated in this study had little effect on how the participants made their final rankings. In fact, the participants in the poorly qualified but chosen condition agreed more with the ECA than the participants in the well qualified but assigned condition.

The third row of the table shows the average number of words the participants used in total during the discussions with the ECAs across conditions. Again, an ANOVA did not yield significant differences. However, a stepwise multiple regression analysis showed that the chosen vs. assigned independent variable predicted the words used over and above the participant's degree of introversion/extroversion; $R^2$ change = .13, $F(1, 56) = 6.7$, $p < .01$. Therefore, if all the participants had the same personality across conditions, then participants in the chosen conditions would discuss the item rankings with the ECA more elaborately than participants in the assigned conditions.

## SESSION TWO

### Study Abroad Program Selection Task

After the trip packing task, each participant performed a second task that involved a hypothetical situation in which

the participant had a friend who was interested in the study abroad programs offered by the university. However, the friend was not able to attend an information session by the international education office, so the friend asked the participant to attend the session and watch the presentation.

Participants met another set of ECAs. Regarding the well qualified vs. poorly qualified variable, in the well qualified conditions the ECAs again looked rather professional and experienced and made a self-introduction such as "I have been working at the office of international education for more than ten years. I saw the expanse of study abroad programs from 20 to more than 50 different programs."

In the poorly qualified conditions, the ECAs were again represented as simple cartoons and made self-introductions such as "I worked at the office of international education for about month. I don't know much about the study abroad programs, but I was asked by the director to give a presentation about the programs because he was on leave."

The participants then watched a presentation made by an ECA which provided some basic orientation information for the study abroad programs. In addition to presenting information through slides, the ECA also engaged the participants through questions and quizzes during the presentation.

After the presentation, the participants were instructed to select a study abroad program for their friend from a number of available programs. They were given an information sheet about their friend's background and preferences. A participant's goal was to narrow down the choices based on all of the background information and the preferences of the participants' friend that were provided.

Because different programs had different requirements and offered different opportunities, the ECA would help the participants choose the program by answering questions and making suggestions. Initially, the ECA would ask the participants some general questions, for example, which courses or subjects their friend would like to study. However, the participants were reminded that, later in the discussion, they should try their best to narrow down the choices by coming up with questions to ask the ECA.

## Measures

### *Subjective Measures*

Similar to session one, four indices were created from the questionnaire items for the study abroad program selection task: likeability, usefulness, satisfaction, and interest.

Likeability was an index made from four adjectives used to characterize the ECA: friendly, annoying, sociable, likable, and one statement: "I would like to complete another online tutorial with the agent" (Cronbach's $\alpha = .88$).

Usefulness was an index made from one adjective used to characterize the ECA: helpful, one adjective used to characterize the presentation by the ECA: informative, three adjectives used to characterize the responses and suggestions made by the ECA: constructive, worthwhile, thoughtful, and one statement: "I felt that it would be difficult to choose the program that best suits my friend without discussing it with the computer agent" ($\alpha = .90$).

Satisfaction of the interaction was an index made from two statements: "I felt comfortable interacting with the computer agent for selecting a study abroad program" and "I enjoyed interacting with the computer agent for selecting a study abroad program" ($\alpha = .93$).

Interest of the task domain was an index made from two adjectives used to characterize the presentation made by the ECA: boring, interesting, and one statement: "I found that studying abroad would be a valuable experience for me after interacting with the computer agent" ($\alpha = .89$).

One last statement was used to measure participant's perceived task performance: "I selected the right program that best suits my friend among all the available study abroad programs."

### *Objective Measure*

This study abroad program selection task was more of an objective task than the trip packing task. It may seem that the study abroad program chosen by a participant would depend heavily on personal subjective preferences. However, the design of the experiment reduced that likelihood. The participants were reminded by the experimenter that they needed to choose the study abroad program that best suited their friend's background information and preferences, which were provided on the information sheet, not on their own personal opinions.

Therefore, the quality of the participant's final choice of a study abroad program can be used as an objective task performance measure. A score was calculated based on how many criteria the participant's choice matched with the provided background information and preferences. For example, because provided background information described their friend as a "bad cook", if the participant decided to choose a study abroad program that offered a meal plan for their friend, then the participant scored one point for this objective measure. The total number of points a participant could possibly have was nine.

## Results

### *Subjective Impressions*

For all analyses we performed full-factorial 2×2 ANCOVAs. Participants' Introversion-Extroversion Scale score, Computer Anxiety Rating Scale score, and their self-rated interests in the study abroad programs from the demographic questionnaire, were used as covariates.

We found significant main effects and interactions on the four indices for the study abroad task similar to the trip packing task (see Figure 3).

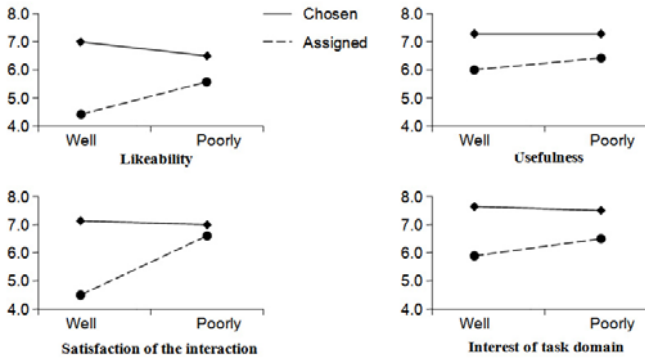


**Figure 3. Mean of Subjective Measures for the Study Abroad Program Selection across Conditions.**

On likeability, there was a significant interaction between well qualified vs. poorly qualified and chosen vs. assigned; $F(1, 56) = 4.2$, $p < .05$. Posthoc pairwise Tukey tests identified that participants in the well qualified & assigned condition viewed the ECA as significantly less likable than participants in the well qualified & chosen and poorly qualified & chosen; $p$'s $< .03$. No other pairwise comparisons reached significance at .05 level.

On usefulness, there was a significant main effect of chosen vs. assigned; $F(1, 56) = 4.0$, $p < .05$. Participants in the choice conditions viewed the ECA significantly more useful than did participants in the assign conditions.

On satisfaction of the interaction, there was a significant interaction between well qualified vs. poorly qualified and chosen vs. assigned; $F(1, 56) = 6.5$, $p < .01$. Posthoc pairwise Tukey tests identified that participants in the well qualified & assigned condition felt significantly less comfortable interacting with the ECA and found the interaction significantly less enjoyable than participants in the other conditions; $p$'s $< .05$. No other pairwise comparisons reached significance at .05 level.

On interest of the task domain, there was a significant main effect of chosen vs. assigned; $F(1, 56) = 4.5$, $p < .04$. Participants in the choice conditions viewed study abroad programs significantly more interesting than participants in the assign conditions did.

### *Task Performance*

Table 3 shows the perceived and actual task performance of the participants. The first row the table shows how well the participants thought they performed the study abroad program selection task across conditions. The analysis showed that there were no perceived task performance differences; none of the pairwise comparisons of the perceived task performance questionnaire item reached significance at .05 level.

The second row shows the number of criteria matched by the participants' final study abroad program selections. The analysis showed that while the trend favored the choice conditions, the differences across the conditions were not significant; none of the pairwise comparison reached significance at .05 level.

| | Well Chosen | Well Assigned | Poorly Chosen | Poorly Assigned |
|---|---|---|---|---|
| Perceived performance | 6.6 | 6.2 | 6.4 | 6.3 |
| Criteria matched | 6.5 | 5.5 | 6.5 | 5.6 |
| Questions helped | 6.3 | 3.7 | 6.4 | 3.9 |

**Table 3. Mean of Objective Task Performance Measures for Study Abroad Program Selection Task across Conditions.**

However, further analysis showed that the participants' final choices were not an accurate indicator of their real objective task performance. Many participants accidentally chose the best matching program at the end by chance without actually taking into account the provided information because they never asked the relevant questions such as whether the program offered a meal plan. Considering that, we further examined the log files of the interactions between the participants and the ECAs. We counted how many questions the participant asked that successfully helped narrow down the program choices based on the provided background information and preferences (duplicated and unrelated questions were discarded). This measure, we believe, more accurately reflects participants' actual task performance. The third row of the table shows that number across the conditions. Participants in the chosen conditions asked more questions that successfully helped narrow down the list of study abroad programs than participants in the assigned conditions; $F(1, 56) = 7.1$, $p < .01$. In conclusion, the participants in the chosen conditions performed better than the participants in the assigned conditions although they were not necessarily aware of that.

## DISCUSSION

Overall this experiment examined the effects of user selection of ECA characteristics on people's perceptions of the ECAs. Quantitative statistical analysis showed that the chosen vs. assigned manipulation had significantly more influence than the well vs. poorly qualified manipulation did. When allowed to choose the characteristics and appearance of an ECA, even if it is just an illusion, participants viewed the ECA as more likable, more trustworthy and more useful. More importantly, this improvement in participants' subjective views of ECAs positively affected their assessment of the task and the quality of interaction and their objective task performance using the ECAs. These results demonstrated that the mere illusion of customization could easily outweigh the cosmetic quality of ECA designs.

In order to understand why and how much this customization effect had on people's perceptions of the

ECAs, we further examined the session transcripts and the interview transcripts; participants were asked a few open-ended questions by the experimenter upon completing the tasks. This qualitative analysis confirmed quantitative findings and provided additional insights.

**Affective Observations**

When the participants were given the chance to choose an ECA to interact with, it did appear that the ECA tended to be more persuasive. One participant put it very plainly during the interview:

> "[In fact] I disagree with her, but I [followed her suggestion and] choose the camera anyway. I chose her because I like her. I like her so I accept her suggestion."

In addition, the ECA evoked more relational behavior from the participants, such as caring about others' feelings, a desire to share, influence and acknowledge, in the chosen conditions. More often in the assigned conditions, the participant tried to avoid discussion and provided very short responses such as:

> "Ok", "No", "I don't know", "Reasonable", "Somewhat", "I imagine so".

In contrast, participants in the chosen conditions tended to put more time and effort into the interactions with the ECAs. As an example, one participant agreed with the ECAs by saying:

> "I agree with that. It makes sense. I would say the translation book probably should not be at the first [place]. Yeah, I forgot about the backpack. [It's an] excellent idea. I can definitely see your point. I did not think of that."

When the ECAs disagreed with the participants, the participants often made concessions or tried to illustrate their points, for example:

> "Yes, I understand that. I was just thinking that maybe because I can find ATM machines there. But it (the ECA's argument about cash being important) makes sense."

Five of the participants in the chosen conditions, when asked by the ECA why they chose a certain item on the list, ended their extensive explanation of the reasoning by asking the ECA for judgment in ways like this:

> "Is that a good answer?"

None of the participants in the assigned conditions exhibited similar responses. They did not seem to care about the ECA's opinions. This might have happened because the participants in the assigned conditions perceived the ECA to be not interested in discussing the item list with them either. In fact, one participant explicitly pointed out that during the interview:

> "Useful? She's like 'This is what I think, like it or not!'"

Most participants in the chosen conditions, in contrast, viewed the ECAs' suggestions positively, for example:

> "I felt the agent was listening to me and decided whether or not my reasoning was good reasons for why I choose the items that I did. I can tell that [her responses were] based on my responses and [she] responded to my responses. It was definitely taking in what I spoke, using it for feedback."

One participant, when asked about possible improvement, actively requested the ECA to be more visible:

> "Can you make the face larger? I like to see her face when discussing the choice with her. It's an interesting experience."

Similar affective effects were found for the study abroad program selection task as well. Participants in the chosen conditions tended to view the task as less difficult and remote and the material as more interesting and important. One participant explicitly commented on that during the interview:

> "Her presentation wasn't as boring as it could have been. It was rather interesting."

In fact, three participants in the chosen conditions actively requested more information that did not necessarily relate to the experimental task by asking questions like this:

> "Could you please give me more information about University of Ulster, Northern Ireland? I am interested in going there myself."

The ECAs they chose appeared to make the participants pay more attention to the topic, which in turn, made them perform better at the task and view the ECA's help as more valuable, like this participant described:

> "[She] sounds educated and knows what she is talking about. She gets the information for me to accomplish the task. [She has] been very helpful. [She] tried very hard to help me to make some decisions."

One participant openly appreciated the ECA's help at the end of the interaction by saying:

> "Thank you for your help. Thank you, Madison (the ECA's name)."

And one participant highly recommended extending this ECA style of interface to other software applications during the interview:

> "It's not the same as you just see something printed on the screen. It's almost like a person talking to you. She seems more like a person. I can imagine that it would be very helpful for learning [other] software."

Again, none of the participants in the assigned conditions had such strong reactions. Although the participants' enthusiasm to perform well for the study abroad program selection task varied across individuals; some participants simply did not really care about the end results; when allowed to choose an ECA to work with, the participants were more motivated to excel. In fact, there were strikingly contrasting descriptions of the ECAs from the participants across conditions.

One participant in the chosen condition during the interview described the ECA as:

> "Friendly, encouraging and respectfully like a *teacher*."

While one participant in the assigned condition portrayed the ECA as:

> "Lecturing like a *teacher* rather than your big sister."

One participant in the chosen condition during the interview described the ECA as:

> "Kind of like a *librarian*, knows her stuff, helpful, attentive, and ready to assist."

While one participant in the assigned condition portrayed the ECA as:

> "Methodical like a *librarian*. More getting down to business than striking you with conversation. Just eliminated the choices."

Finally, we found that performance of the ECA could evoke strong affective reactions from the participants. In an earlier study we conducted that also used the trip packing task [4], an unintentional lag happened after a participant argued her disagreement with the ECA's suggestions. Later, during the interview, when asked about the ECA's personality, the participant commented:

> "I found the ECA to be too sensitive, maybe. He seemed almost irritated when I didn't agree with him and just stood there."

In the present study, mistakes also were not merely interpreted as innocent program bugs. Instead, they had unexpected consequences. As an example, one participant complained during the interview:

> "I thought the computer lied to me once. She showed me that Technical University of Munich, Germany was among the list [of programs] that had no GPA requirement. But later she told me that my friend does not qualify because her GPA is too low."

It seems that one little mistake can have drastic impact on how the ECA is perceived. One participant highlighted this issue during the interview:

> "Once she said "that was very good argument." I wasn't really making any argument, I was agreeing with her by talking about it. When she said that was very good argument, [it] kind of reminded me [that] it was a computer program type of agent. Every other time it seemed that I was actually speaking to someone face to face. That was one time that I was like 'wow, that *was* a computer program.'"

#### Individual Preferences

One important piece of information gathered during the interview is why the participants in the chosen conditions chose the particular ECA from the set of three possibilities. Participants reported many different rationales. Some simply liked the appearance of the ECA:

> "Her appearance seems like she was more presentable."

> "I thought she was more lively."

Some thought the ECA they chose had a better voice although in reality all the ECAs used the same speech synthesizer with the same set of parameters:

> "She is the easiest to follow. [Her voice] is a little slower and smoother."

> "I like the sound of her voice. Her speech is little easier to understand."

Consistent with findings from previous studies [9], individual participants varied greatly in their preferences of ECA characteristics. The fact that no particular ECA was favored by a majority of the participants and the fact that the participants made different selections on the preference sheet has already demonstrated that people have diverse, sometimes contradictory, views on how an ECA should look, sound and behave. In all likelihood, particular people's background, experience, and personality made them more or less comfortable with the different ECAs.

For example, one participant explained why she chose the ECA:

> "[I chose her] because she seems to be the nicest person and easiest to talk to. I like the fact that she is more of a *casual* traveler. I was more ease with her. She was more amiable."

Whereas, another participant who chose a different ECA had the exact opposite rationale:

> "[I chose her because] she is very knowledgeable, very *professional*. And [I believe she would] be able to help me the best."

One participant chose a younger looking ECA because she felt:

> "She (the ECA) is the one I can most relate to because of her *age*."

Whereas, another participant thought differently:

> "I thought she was very experienced in traveling. The top agent, she seemed pretty experienced but she was much *younger*, may not be as responsible in the traveling situation as Debbie seemed she would be. That's why I choose Debbie."

Quite often, it is the personal relationship the participants found with the ECA that influenced their decisions. One participant's response during the interview, we believe, highlighted this issue:

> "I didn't enjoy France, but I enjoyed Spain. I choose Christy merely because she has been to Spain."

The result from this study suggests that the effect of user choice and customization can out-weigh the perceived quality effect of the ECAs. In this study, there was actually not much user customization, except for the fact that in the chosen conditions the participants simply had choices. The findings suggest that it might not be worthwhile to spend great effort building the *perfect* ECA for all users. Instead,

providing a customization UI for ECAs that is flexible and easy to use might be a better approach.

In fact, it seems almost impossible to design an ECA that would suit all users. One participant in the study did not like the ECA simply because the performance of the ECA was too good!

> "She doesn't hesitate, pretty fast on everything, but maybe too fast, too serious and too logical. A database query [system] would have achieved the same thing. Humans make mistakes."

Finally, a surprising result from the subjective impression statistical analysis provided further evidence showing that people's perception of an ECA may deviate greatly from the models that ECA designers intend to portray. For the study abroad program selection task, the well qualified & assigned ECA was the least likable and enjoyable among the four conditions, not the poorly qualified & assigned ECA. The nature of the task might have played a role here as reported by other researchers [6, 7]. The study abroad task is more of an objective task (find the best program that matches given information) whereas the trip packing is more of a subjective task (there is no best ranking). During the interview, the participants often commented that they did not enjoy the assigned formal instructor-looking ECAs for the study abroad task. The participants who interacted with the assigned cartoonish ECAs, on the other hand, seemed to be more comfortable.

## CONCLUSIONS

There have been noteworthy failures in ECAs, such as the Microsoft's Office Assistant "Clippy". Anecdotally, we have observed that people who did change the MS Office Assistant to a dog or a cat appeared to have a more positive view of the agent and continued using it, which seems to align with findings of our study.

This experiment demonstrated that users build up impressions of an ECA by extrapolating various cues from its utterances and appearance, and those impressions are judged and evaluated against the person's preferences, beliefs and abilities and, which in turn, determine their attitudes and behaviors when interacting with the ECA. User customization can have a significant positive effect on people's perception of an ECA. The customization process, not just the customization end results, had a strong influence on people's perception of the ECAs.

Furthermore, the interaction transcripts and performance data reveal that customization could have a positive effect on people's behavior. When allowed to choose an ECA, people were more interested in the topic, more active with the interaction, more likely to take the agent's advice, and more motivated to succeed. This motivation effect can be very beneficial in application domains such as learning [5].

Finally, the study results demonstrated that user expectations about and perceptions of the interaction between themselves and an ECA depend very much on the individual's preconceived notions and preferences of various ECA characteristics and might deviate greatly from the models that ECA designers intend to portray. Much work has been done on getting characteristics just right. Our research shows that it is partially misdirected. Instead, giving people choices and allowing customization might be a better approach. History has taught us that artifacts are designed that people then use in ways that are quite different from what the designer expected. As the use of ECAs in software increases, it is essential that designers of these ECAs work with people's preferences to customize the right perception into designs. As UI designers, we need to allow people to choose from ECA interfaces, or even non-ECA interfaces, with a variety of characteristics that match their preferences, not assign users with ECAs that we think would work best for them.

## REFERENCES


1. Bickmore, T.W. Relational agents: Effecting change through human-computer relationships, Massachusetts Institute of Technology, 2003.
2. Brosnan, M.J. The impact of computer anxiety and self-efficacy upon performance. *Journal of Computer Assisted Learning*, *14* (3). 223-234.
3. Cassell, J. Embodied conversational interface agents. *Communications of the ACM*, *43* (4). 70-78.
4. Catrambone, R., Stasko, J. and Xiao, J., Anthropomorphic Agents as a User Interface Paradigm: Experimental Findings and a Framework for Research. In *Proceedings of CogSci* 2002, 166-171.
5. Cordova, D. and Lepper, M. Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. *Journal of Educational Psychology*, *88* (4). 715-730.
6. Cowell, A.J. and Stanney, K.M., Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents. In *Proceedings of IVA* 2003, 301-309.
7. McBreen, H., Anderson, J. and Jack, M., Evaluating 3D Embodied Conversational Agents in Contrasting VRML Retail Applications. In *Proceedings of Workshop of Multi-Modal Communication and Context in Embodied Agents,* 2001, 83-88.
8. Nass, C. and Lee, K.M., Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Proceedings of CHI 2000*, ACM Press, 329-336.
9. Xiao, J., Stasko, J. and Catrambone, R., An empirical study of the effect of agent competence on user performance and perception. In *Proceedings of AAMAS 2004*, IEEE Computer Society, 178-185.