"Wow, You Are Terrible at This!" - An Intercultural Study on Virtual Agents Giving Mixed Feedback

Isaac Wang
wangi@ufl.edu
University of Florida
Gainesville, FL, USA

Lea Buchweitz
lea.buchweitz@hs-offenburg.de
Offenburg University
Offenburg, Germany

Jesse Smith
jd.smith@ufl.edu
University of Florida
Gainesville, FL, USA

Lara-Sophie Bornholdt
Jonas Grund
lbornhol@stud.hs-offenburg.de
jonas.grund@hs-offenburg.de
Offenburg University
Offenburg, Germany

Jaime Ruiz
jaime.ruiz@ufl.edu
University of Florida
Gainesville, FL, USA

Oliver Korn
oliver.korn@hs-offenburg.de
Offenburg University
Offenburg, Germany

ABSTRACT

While the effects of virtual agents in terms of likeability, uncanniness, etc. are well explored, it is unclear how their appearance and the feedback they give affects people's reactions. Is critical feedback from an agent embodied as a mouse or a robot taken less serious than from a human agent? In an intercultural study with 120 participants from Germany and the US, participants had to find hidden objects in a game and received feedback on their performance by virtual agents with different appearances. As some levels were designed to be unsolvable, critical feedback was unavoidable. We hypothesized that feedback would be taken more serious, the more human the agent looked. Also, we expected the subjects from the US to react more sensitively to criticism. Surprisingly, our results showed that the agents' appearance did not significantly change the participants' perception. Also, while we found highly significant differences in inspirational and motivational effects as well as in perceived task load between the two cultures, the reactions to criticism were contrary to expectations based on established cultural models. This work improves our understanding on how affective virtual agents are to be designed, both with respect to culture and to dialogue strategies.

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; • **Social and professional topics** → **Cultural characteristics**; • **Human-centered computing** → *User studies.*

KEYWORDS

Conversational agents, virtual agents, culture, agent feedback

IVA '20, October 19–23, 2020, Virtual Event, Scotland, UK
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423887

ACM Reference Format:
Isaac Wang, Lea Buchweitz, Jesse Smith, Lara-Sophie Bornholdt, Jonas Grund, Jaime Ruiz, and Oliver Korn. 2020. "Wow, You Are Terrible at This!" - An Intercultural Study on Virtual Agents Giving Mixed Feedback. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland, UK*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423887

1 INTRODUCTION

A key use for agents is to assist, and provide feedback and motivation. Virtual agents or, more specifically, embodied conversational agents, extend the capabilities of dialogue systems by providing the system with a visual embodiment [10, 32, 41]. Traditional dialogue systems already enable natural user interaction by allowing users to speak and converse with a computer through natural discourse. Virtual agents take that core concept a step further by providing an embodiment (examples in Fig. 1), which allows an agent to take on a more social role during the interaction: users can now communicate with "something" they can relate to and ideally with an entity which is familiar or even likeable. Due to their ability to assist and socialize on a more human level, agents have been employed in widespread applications, from teaching [33, 39] to healthcare [13, 18].

There is considerable research on the effects of agent appearance and agent perceptions in terms of likeability and uncanniness [8, 34, 43]. However, it is unclear how an agent's appearance affects how people interact with it: is feedback from a mouse or a robot taken less serious than feedback from a human agent? How does this change if the feedback is not supportive but critical? Additionally, there is little work examining how cultural background affects the perception of agents and their feedback. For example, some cultures are very careful when framing criticism (e.g. the United States), whereas in other cultures negative feedback is a much more regular and accepted part of communication (e.g. Germany).

We designed a study with a hidden-object game, in which an agent provides feedback to the user, to understand how agent appearance and participant cultural background affect the perceptions of agent feedback. Our work contributes to a new understanding on how users' cultural backgrounds affect agent perceptions and thus informs the design of future motivational agents.

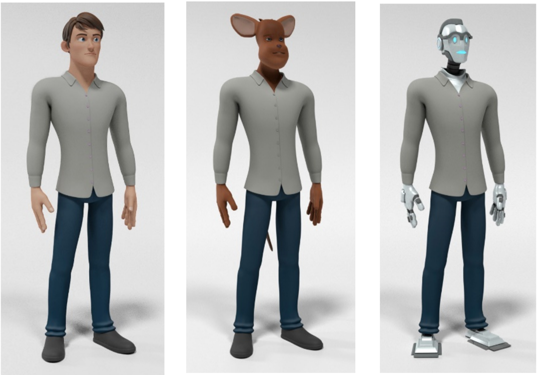

Figure 1: The virtual assistants presented to the participants.

2 RELATED WORK

Embodied Conversational Agents [10], or ECAs, are systems that pair the natural speech interface of a dialogue system with a visual embodiment or avatar. Embodiment personifies a system [26], allowing it to take on a social role when interacting with users [32, 41]. These virtual agents have been employed in different applications, ranging from teaching [33, 39] to healthcare [13, 18].

In particular, some of the agents in teaching and coaching settings try to keep users motivated on a task by providing active feedback to the user. Motivational agents are primarily used in instructional scenarios, due to their influence on learning outcomes [6, 24]. These agents provide both positive and negative feedback to guide users to success [27, 31, 40]. Due to their ability to function on a social and affective level, feedback-giving agents are able to increase intrinsic motivation in users [31]; likewise, social robots have also been shown to evoke affective reactions in users [9], which is critical when delivering praise and criticism [30, 46].

For our research, we are interested in understanding how the appearance of an agent affects how users react to and perceive the agent's feedback, and how the cultural background of a user can impact those perceptions. We focus our review of related work on two areas: the impact of appearance in ECAs, and cultural factors involved in human-agent interaction.

2.1 Effects of Agent Appearance

The appearance of an agent can affect users in various ways. When comparing a teddy bear agent, a human-like agent, and an agent with a box as a head, Bailenson et al. [4] showed that the teddy bear and human-like agents were rated as more likeable than agent with box-like features. In addition, for agents in augmented reality applications, Wang et al. [43] showed how agents that were perceived to be more human were more likeable and personable than one that appeared more like a machine.

The anthropomorphism of an agent also has a large effect on users [5]. Parise et al. [34] showed people two different agents: one with a human-like appearance and one with a dog-like appearance. They found that people were more likely to cooperate with the human-like agent as opposed to the dog-like one. Bergmann et al. [8] note that, compared to human-like agents, robot-like agents must exhibit more consistent human-like behaviors in order to maintain the same level of perceived warmth. We seek to understand how these differing levels of human-likeness can affect how users perceive the feedback given by a motivational agent and compare how users with different cultural backgrounds react.

2.2 Cultural Factors in Agents

Researchers have also focused on creating agents that model and portray different cultures [2, 29, 36, 37], follow specific social rules [12, 25], or adapt to the culture of the user [35, 38]. For example, Mascarenhas et al. [29] created groups of agents with differing behaviors based on their own cultural standards. When evaluating these agents, they found that people were able to discern the behaviors between groups and ascribe the differences to the agents' cultures. In line with these findings, Rosis et al. [35] suggest how an agent must be adaptable to different cultures and contexts in order to communicate effectively with different users. In our case, we focus not on creating culturally-adaptive agents, but take a step back and understand how a user's cultural background affects how they perceive an agent.

Other researchers have studied how users' cultural background affects the way they respond to an agent. Endrass et al. [14] showed how users prefer to interact with an agent matching their own culture, in terms of both spoken and nonverbal communication. Additionally, Ishioh and Koda [23] and Isbister et al. [22] showed how a user's cultural background influences how comfortable they are with an agent's nonverbal behaviors (such as proximity and self-touching). In our work, we aim to understand how a user's cultural background affects the affective perception of an agent and its feedback.

2.3 Hofstede's Model of Cultural Dimensions

Geert Hofstede defined culture as "the collective programming of the mind that distinguishes the members of one group or category of people from others [20]." He argued for the possibility to compare cultures if we imagine the characteristics of individuals vary according to the bell curve, and the shift of the bell curve when one moves from one society to the other describes cultural differences [20, 21].

In his initial work, Hofstede performed ecological factor analysis on survey data collected from IBM employees in over 50 countries to propose 4 dimensions of culture [19]. Follow up work by him and his colleagues added two additional dimensions resulting in six dimensions [21], often referred to as "Hofstede's Model of Cultural Dimensions." The model represents each international culture as numeric ratings along the six dimensions [21], ranging from individualism to restraint. The empirical nature of these dimensions allows comparisons of each culture's values and attitudes.

3 THE GAME AND THE AGENTS

We created a "find-the-hidden-object game" in which users interact with a virtual agent that gives hints and provides feedback on their performance. The task was inspired by prior work [16, 44], which asked people to find objects while engaging with an agent that knew where the objects were located. To understand how agent appearance and user cultural background affect the perception of the agent's feedback, we evaluated participants' emotional reactions.

3.1 Hidden-Object Game

The hidden-object game was developed with the Construct 2 software and designed to run on a touchscreen. The game had ten one-minute levels of increasing difficulty; in each level the player had to find a hidden purple star, covered by colored geometrical shapes (i.e., circles, rectangles, triangles and squares in green, blue, red, or yellow). The player was required to drag the shapes out of the way to reveal the star, then tap on it to complete the level. To ensure that all participants experienced the same amount of negative feedback by the virtual agent, four of the ten levels (six, eight, nine, and ten) were designed to be unsolvable (the star was hidden below a very high number of shapes). However, participants did not know that any of the levels were unsolvable.

During each level, the player could ask the virtual agent for hints on where the star was hidden and the agent would verbally respond with additional information. Participants could ask for information on (i) the star's location (top vs. bottom and left vs. right), (ii) the color of the object covering the star, or (iii) the shape of the object covering it. In order to prevent players from constantly asking for hints, we restricted the number of hints to three per level. In addition, the first hint was available five seconds after starting, with a 15 second "cool-down" between remaining hints.

3.2 Virtual Agents

The agent's appearance can be seen as an independent variable with three values: Human, Mouse, and Robot. Our study was designed as a 3x2 (agent x interaction style) between subjects study; thus, each participant only experienced one agent appearance and interacted with it in only one of the two modalities. For the three appearances, the agents share an anthropomorphic figure, as depicted in Figure 1. In addition, we designed two different ways to interact with the agent. In one condition, participants interacted with the agent by tapping buttons on the game interface. In the other condition, the participants interacted verbally with the agent. All other aspects of the agents (voice, gesture, etc.) were controlled between conditions. For brevity, we will not report on the two interaction styles in this paper, but instead focus on the cultural aspects of our study.

At the end of each level, the agent would give positive or negative feedback based on the player's performance. The feedback was presented the same way regardless of agent or interaction style. As the player completed or failed multiple levels, the agent would increase the intensity of praise or criticism. The first stage of feedback was provided after the first level was won or lost, the second stage occurred when the participant had won or lost two levels in a row, and the third stage was presented when the participant won or lost three levels in a row. After the third win/loss in a row, the feedback reset to the first stage. The praise feedback phrases (listed in order of increasing intensity) were: "You did it," "You found that pretty quickly," and "Wow, you are really good at this." The criticism phrases (in increasing intensity) were: "Time's up," "You really took a long time," and "Wow, you are terrible at this."

4 METHODOLOGY

Our study had two primary goals. First, we wanted to understand if different agents, in spite of similar anthropomorphic features, will effect participants' reactions and perceptions when receiving both positive and negative feedback regarding their task performance. Second, we wanted to explore how cultural background might effect the participants' responses to negative and positive feedback. In this section, we present our hypotheses.

4.1 Hypotheses

H1: The level of an agent's human likeness effects participants' response.

Prior work has shown that both the appearance of an agent [4, 43] and the level of anthropomorphism [7, 8, 11, 34] affect users' perceptions. In our study, three different agents (Human, Mouse, and Robot) share a similar level of anthropomorphism (i.e., a human-like body) but differ in the appearance of their head, hands, and feet. Based on prior work, we expect to see differences in the perception of the human agent in respect to the other two agents. We also want to explore how the agents' appearance may elicit different reactions from participants when receiving both positive and negative feedback. More specifically, we expect participants to have stronger negative reactions to an agent with more human-like features. As users are more sensitive to human-like agents [8], the human agent's negative feedback should be perceived more negatively than corresponding feedback from the robot and the mouse. Prior work also suggests that users are more sensitive to robots [8]; thus, we hypothesize that participants would react less negatively to criticism from the mouse. Therefore, we split H1 into the following three sub-hypotheses:

- H1.1: Participants will rate agents more negatively the more human-like they look.
- H1.2: The participants' positive affective response to positive agent feedback will increase with the agent's level of human-likeness.
- H1.3: The participants' negative affective response to negative agent feedback will increase with the agent's level of human-likeness.

H2: Participants with cultural backgrounds from Germany versus the US will have different affective responses to agent feedback. Hofestede's Model of Cultural Dimensions [21] describes several differences between Germany and the United States. As such, we expect our participants to have different affective responses to both positive and negative agent feedback given their cultural background. We hypothesize that the differences in Uncertainty Avoidance, Long Term Orientation, and Indulgence dimensions [21] will result in persons with a US background having stronger positive reactions to positive feedback and more negative reactions to negative feedback than their counterparts with a German background.

Hofestede states that cultures that score high on the Indulgence dimension are categorized by a people who are more likely to remember positive emotions than those in a restrained culture [20]. Thus, we hypothesize that participants with a US background will remember the feedback more than their German counterparts resulting in higher affective ratings to positive agent feedback (H2.1).

Cultures that score low on Long Term Orientation tend to attribute success and failure more to luck than cultures which score high, who attribute success and failure to effort [20]. This suggests that participants with a background from the US will externalize

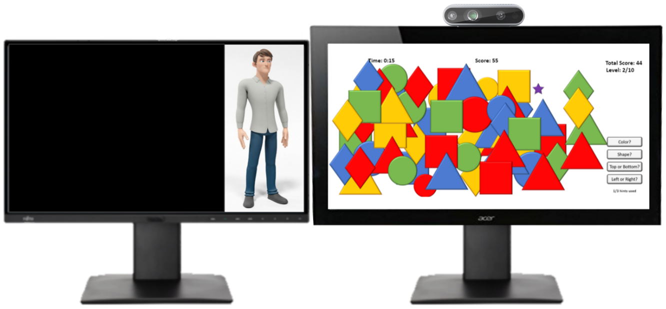

Figure 2: Apparatus used for the experiment.

their personal performance compared to German participants, leading to negative reactions when criticized on a task they perceive as having little control over. Additionally, we would expect that, due to high Individualism, low Power Distance, and low Uncertainty Avoidance, US participants would act more negatively towards an agent that criticizes them because they don't believe the agent is in a position of authority to provide such feedback (H2.2). As persons with a German background score higher in the Long Term Orientation dimensions and attributing success and failure more to personal effort rather than to luck, they will perceive tasks to require higher cognitive and physical demand (H2.3). We summarize these sub-hypotheses below:

- H2.1: Participants with a US background will respond more positively to positive agent feedback than those with a German background.
- H2.2: Participants with a US background will respond more negatively to negative agent feedback than those with a German background.
- H2.3: Participants with a German background will rate the task to have higher physical and mental demands than those with a US background.

4.2 Participants

We recruited 120 participants across two institutions: 60 participants were recruited from a German university (32 male, 28 female, mean age = 22.90, SD = 4.56), the other half from an US university (34 male, 26 female, mean age = 21.68, SD = 2.45). Overall, participants were adults between the ages of 18 and 46 (M=22.29, SD = 3.71) consisting of 66 males and 54 females.

4.3 Apparatus

The study setup consisted of two monitors standing right beside each other (see Figure 2). Whereas the game was played on a touch screen monitor (Acer UM.HT2AA.002 Touch, 27-inch, Full HD), the virtual agent was displayed on the right border of the adjacent monitor of the same size on the left of the touch screen monitor (FUJITSU Display P27-8 TS Pro, 27-inch, Full HD).

On the touch screen monitor, an Intel RealSense camera was mounted to record participants' faces. The face recordings were used to track fixations on the left-hand monitor, i.e., on the virtual agent instead of the game. For scene recordings, another video camera was placed diagonally on the left behind participants to capture participants' reactions and gain additional observations.

4.4 Procedure

At the start of the study, the researcher described the rules of the game to each participant: that they would need to drag and move the shapes in each level to find the star, and that the agent knows where the star is and could provide hints. The participant was randomly assigned to one of the three agent conditions (Human, Mouse, or Robot) which dictated the agent they would be interacting with and the interaction condition (interface button or speech). The participant then completed all 10 levels of the game in the same order. After completing the game, each participant filled out a questionnaire describing his or her experience with the task, the agent, and the feedback they received (see following section). At the end of the study, each participant was compensated ($20 or €20) for their time.

4.5 Post-Study Questionnaire

The questionnaire consisted of five different sections, evaluating different dimensions of participants' experience with the agent:

The first section comprised questions regarding participant's demographic information, including ethnicity, gender and age, as well as additional information on handedness and visual or hearing impairments. The second part of the questionnaire asked the participants to rate the difficulty of the task, using the NASA-TLX survey [17] for measuring task load across six scales. In the third section, participants were asked to rate six different qualities of the virtual agent on five-point Likert scales. These questions included ratings for if the agent was helpful, personal, trustworthy, appropriate, likeable, and if the user would be willing to interact with the agent again (continued use). These questions were adapted from prior work on embodied agents [3, 43] and which we refer to as Agent Rating Questionnaire (ARQ) in the rest of this paper.

Sections four and five of the questionnaire assessed the positive and negative affective reactions of the participants towards the feedback given by the virtual agent. Both sections comprised the questions from the International Positive and Negative Affect Schedule Short-Form (I-PANAS-SF) [42] to assess positive and negative affective reactions [42] on five-point Likert scales. Lastly, participants were asked open-ended questions on what they liked most/least about the agent and their thoughts on the feedback.

5 RESULTS

In this section we present our analysis of responses to the post-study questionnaire. When conducting an analysis of variance (ANOVA), we first ran a Shapiro-Wilks test for normality. If the distribution was not normal, we applied an Aligned Rank Transform (ART) [45] to the data. As mentioned earlier, we do not report differences in interaction styles (interface or speech), but instead leave the analysis for future work. Thus, we do not consider interaction style an independent variable in our analysis for the results presented.

5.1 User Responses on Agent Appearance (H1)

Our first hypothesis and sub-hypotheses focus on the effects of agent appearance on how participants perceive the agent, and on

Table 1: Means and standard deviations of I-PANAS-SF dimensions of both encouragement and criticism by agent.

Affective Response Measures to Encouragement

Agent	Active	Alert	Attentive	Determined	Inspired	Afraid	Ashamed	Hostile	Nervous	Upset
Human	3.61 (0.95)	2.95 (1.07)	3.34 (0.96)	3.59 (1.00)	3.39 (1.02)	1.51 (0.98)	1.68 (1.04)	1.44 (0.90)	2.10 (1.20)	1.76 (0.97)
Mouse	3.66 (0.63)	3.05 (1.21)	3.55 (1.01)	3.84 (0.97)	3.24 (0.94)	1.50 (0.83)	1.55 (0.92)	1.55 (0.89)	1.89 (1.05)	1.76 (0.97)
Robot	3.41 (1.09)	3.11 (1.05)	3.19 (1.05)	3.57 (0.99)	3.41 (1.01)	1.30 (0.70)	1.24 (0.76)	1.38 (0.89)	1.68 (1.08)	1.46 (0.80)

Affective Response Measures to Criticism

Agent	Active	Alert	Attentive	Determined	Inspired	Afraid	Ashamed	Hostile	Nervous	Upset
Human	2.85 (1.01)	3.22 (0.94)	2.76 (0.94)	3.29 (1.15)	2.68 (1.21)	1.83 (1.16)	2.61 (1.20)	2.88 (1.23)	2.49 (1.08)	2.85 (1.24)
Mouse	2.84 (1.03)	3.16 (1.08)	3.03 (1.08)	2.97 (1.24)	2.79 (1.09)	1.71 (0.98)	2.39 (1.13)	2.89 (1.39)	2.47 (1.29)	2.74 (1.13)
Robot	2.59 (1.21)	2.87 (1.30)	2.62 (1.18)	2.95 (1.38)	2.31 (1.17)	1.44 (0.79)	2.18 (1.37)	2.38 (1.33)	2.21 (1.30)	2.36 (1.04)

the participants’ affective reactions towards the agent. For user perception, we examined the results of the Agent Ratings Questionaire (ARQ) of each agent. In Figure 3, we see that the agents were rated similarly: an ANOVA revealed no significant difference of agent on user perception ($F_{2,116} = 0.75$, n.s.). We also analysed each perception metric individually. Again, there was no significant effect of the agent. Therefore, we reject H1.1, that user perception would be more negative with increased human-likeness.

To test the effects of feedback on participants affective reactions (H1.2 and H1.3), we analyzed the scores of the I-PANAS-SF questionnaires. We examined each type of feedback the agent gave (encouragement vs. criticism) and for each feedback type we derived an aggregate score for "Positive Affect" and "Negative Affect" by totaling the dimensions corresponding to each affect (e.g. Active, Alert, Attentive, Determined, and Inspired for positive affect) [42]. Means and standard deviations of the individual dimensions on the I-PANAS-SF scale are shown in Table 1.

Four participants did not respond to the questions about the *agents’ encouragement*, so we conducted our analysis on the remaining 116 participants. For affective reactions towards the agents’ encouragement, an ANOVA found no significant effect of agent appearance on positive affective reactions ($F_{2,113} = 0.15$, n.s.), however, a significant effect of agent appearance was found on negative affective reactions ($F_{2,113} = 3.21$, $p < 0.05$). Post-hoc analysis using Tukey correction showed that negative affective reactions were higher for the Human agent than for the Robot agent ($p < .05$). No differences were found between the Mouse agent and the Human or Robot agents. An analysis of the individual dimensions of the I-PANAS-SF showed no significant effects of agents on any of the individual dimensions. Since no significant effect of agent appearance on positive affective reactions was observed we reject H1.2.

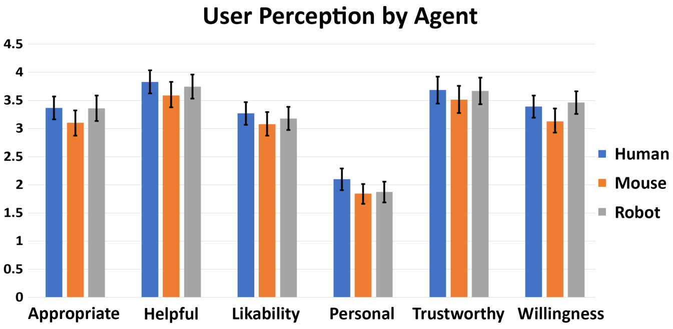

Figure 3: Average user perception metric by agent. Error bars represent 95% confidence intervals.

Next, we analyzed the affective reactions regarding the *agents’ criticism.* Two participants did not respond to the questions for agent criticism, so we conducted our analysis on the responses from the remaining 118 participants. An ANOVA identified no significant effect of an agent, both on positive affective reactions ($F_{2,115} = 1.24$, n.s.) and on negative affective reactions ($F_{2,115} = 2.27$, n.s.). In addition, an analysis on the I-PANAS-SF individual dimensions showed no significant effects of agents on the dimensions. As the participants’ negative reactions to negative feedback were consistent among the three agents, we had to reject H1.3.

For completeness, we also examined the responses of the NASA-TLX survey although this data was not needed to address the H1 hypothesis and we did not expect any differences in perceived task load by agents, as the tasks and the behaviors of the agents were identical. Accordingly, an ANOVA found no significant effect of the agents on the NASA-TLX measures ($F_{2,117} = 0.13$, n.s.).

5.2 Affective Responses across Cultures (H2)

Our second hypothesis and sub-hypotheses focus on differences in participants’ affective responses based on the cultural background. To analyze affective responses, we again examined the results of the I-PANAS-SF. Results are shown in Table 2.

Four participants did not respond to the questions about the agent’s encouragement leaving 116 participants for this analysis. The distributions were found to be non-normal, so we applied the ART transform before analysis. For negative affective reactions towards the agent’s encouragement, an ANOVA revealed a significant effect of culture ($F_{1,114} = 8.87, p < 0.01$), with Germans having greater negative reactions towards positive feedback from the agent. No significant main effect of cultural background on positive affective reactions was found ($F_{1,114} = 1.00$, n.s.). An analysis of the individual dimensions of the I-PANAS-SF revealed a significant main effect of cultural background on the *upset* dimension ($F_{1,114} = 13.3$, $p < 0.001$) and *nervous* dimension ($F_{1,114} = 6.22$, p <

Table 2: Means and standard deviations of I-PANAS-SF dimensions to both encouragement and criticism by study location. Significance denoted by: p < 0.05 (*), < 0.01 (**), and < 0.001 (***).

Affective Response Measures to Encouragement

Location	Active	Alert	Attentive	Determined	Inspired	Afraid	Ashamed	Hostile	Nervous*	Upset***
Germany	3.48 (0.89)	3.14 (0.96)	3.30 (0.93)	3.77 (0.95)	3.36 (1.12)	1.50 (0.97)	1.59 (1.06)	1.46 (0.93)	**2.13 (1.18)**	**1.95 (0.98)**
U.S.	3.63 (0.92)	2.93 (1.22)	3.42 (1.08)	3.57 (1.01)	3.33 (0.86)	1.38 (0.72)	1.42 (0.79)	1.45 (0.85)	**1.68 (1.05)**	**1.40 (0.79)**

Affective Response Measures to Criticism

Location	Active	Alert	Attentive**	Determined*	Inspired**	Afraid	Ashamed	Hostile	Nervous*	Upset
Germany	2.67 (1.18)	2.95 (1.05)	**2.53 (1.03)**	**2.83 (1.24)**	**2.88 (1.17)**	1.79 (1.10)	2.50 (1.26)	2.81 (1.38)	**2.67 (1.23)**	2.50 (1.14)
U.S.	2.85 (0.99)	3.22 (1.06)	**3.05 (1.06)**	**3.32 (1.23)**	**2.32 (1.11)**	1.53 (0.87)	2.30 (1.23)	2.63 (1.28)	**2.12 (1.15)**	2.80 (1.15)

0.05). In both cases, participant with a German background rated the terms higher than their US counterparts. Though our results indicate that Germans tend to have more negative reactions towards compliments than US participants, we must reject H2.1 since there were no differences in how each culture positively reacted to positive feedback.

We next examined the participants' reactions to criticism from the virtual agent (H2.2). As two participants did not respond to the questions about the agent's criticisms, the analysis is based on the remaining 118 participants. An ANOVA shows no significant effect of cultural background on either positive affective reactions ($F_{1,116}$ = 0.71, n.s.) or negative affective reactions ($F_{1,116}$ = 1.86, n.s.) to criticism from the virtual agent. When examining the individual terms of the I-PANAS-SF, we noted significant differences based on cultural background: an ANOVA identified a main effect on the *determined* ($F_{1,116}$ = 5.19, p < 0.05), *inspired* ($F_{1,116}$ = 7.40, p < 0.01), *attentive* ($F_{1,116}$ = 7.48, p < 0.01), and *nervous* ($F_{1,116}$ = 6.34, p < 0.05) dimensions. Participants with a German background reported higher ratings for the *inspired* and *nervous* dimensions, but participants with an US background reported higher ratings for the *attentive* and *determined* dimensions. With these results, we must reject H2.2 since US participants reported no significant negative reactions to the negative feedback of the virtual agent.

Finally, we examined the results of the NASA-TLX to test our hypothesis that perceived task load would be effected by cultural background (H2.3). An ANOVA revealed a significant effect on the weighted TLX scores ($F_{1,118}$ = 4.06, p < 0.05). When conducting an analysis on the raw individual scale (after applying an ART transform since the scales were found to be non-normal), we saw significant effects of cultural background on the *mental demand* ($F_{1,118} = 9.75, p < 0.01$) and *physical demand* ($F_{1,118} = 8.47, p < 0.01$) scales (Figure 4). Participants from Germany reported significantly higher *mental demand* ($M = 27.0, SD = 17.1$) compared to participants from the US ($M = 18.7, SD = 14.5$), while participants from the US reported higher *physical demand* than participants from Germany ($M = 61.7, SD = 26.8$ and $M = 48.3, SD = 25.4$ respectively). No other significant effects were found for the individual scales. Thus, our results support hypothesis H2.3, but not necessarily in the manner we expected.

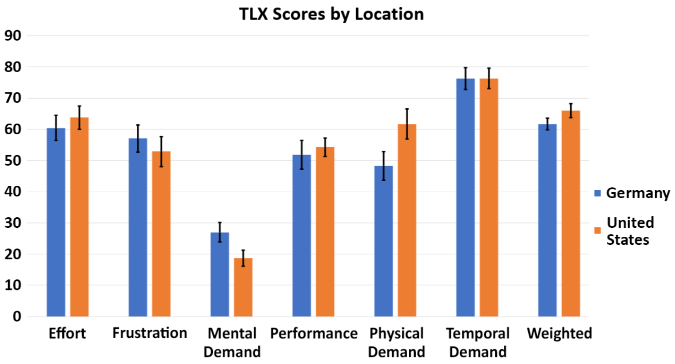

Figure 4: Average NASA-TLX scores by location. Error bars represent 95% confidence intervals.

6 DISCUSSION

6.1 Agent Appearance and User Perceptions

Based on the user study, we had to reject H1. Contrary to prior work, the results showed the agents' appearance did not change the participants perception of the agent (H1.1). However, we attribute this to the fact that agents' levels of anthropomorphism and appearance may have been too similar to cause different perceptions. Recall that we purposefully created agents with similar levels of anthropomorphism, only changing the appearance of the head and hands, but retaining the same high quality facial expressions and features. The time sensitive nature of the task and the placement of the agent on the second screen may also have relegated the agent to an "assistant" role, instead of being the focus of participants' attention.

While participants' perceptions of the agents were not different, we did see that participants had a more negative reaction towards encouragement from the human than the robot (with mouse falling in the middle). This contradicts our initial hypothesis that participants would have the strongest positive reactions to the human agent when receiving positive feedback. Recall that Bergmann et al.'s study [4] found that people initially react positively to robot agents only to see their perceptions decrease over time. Whereas, the human agent in their study had more stable ratings. Given that participants were first presented with positive feedback, participant ratings would have been captured during their first interactions with the robot, possibly causing higher affective responses compared to the other agents. Future work should further explore these perceptions over time when regarding positive feedback, especially since it is the common type of feedback presented to users when interacting with virtual agents.

6.2 Cultural Factors in Perceptions of Agents

The other focus of our study was to identify cultural differences in the perception and affective responses of agent feedback (H2). First, we saw that overall perceptions of task load differed by culture: participants from the US had significantly higher TLX scores overall, higher Physical Demand, and lower Mental Demand. This supports our hypothesis (H2.3) that we would see differences in perceived task load based on Hofestede's Cultural Dimensions where cultures considered to focus on a long term orientation (such as Germany) attribute success and failure to one's effort opposed to luck. However, when identifying cultural differences in perceptions and affective responses of agent feedback, surprisingly participants from the US had higher affective responses for the *attentive* and *determined* dimensions and lower response for the *inspired* and *nervous* dimensions when receiving criticism. When we examine the open-ended feedback left by participants, US participants stated a desire to prove the agent wrong, possibly explaining a heightened sense of attentiveness and determination. For example P12 stated, "I was not expecting it, and felt like I had to prove it wrong!" This directly contradicts our hypothesis based on Hofstede's principles where US participants would externalize the negative feedback causing a more negative reaction.

From our results, it is clear that we cannot assume an agent's feedback will be appropriate for all cultures. In addition, while task load was sufficiently predicted by cultural models, we find that relying solely on insights based on existing models may not be sufficient for designing agent feedback. Based on Hofstede's Cultural Dimensions, we expected a stronger negative reaction by US to criticism. However, this was not the case. It turns out that participants' reactions to the feedback was not obvious and more nuanced than the model could predict.

Prior work has predominantly relied on generalized cultural models (e.g. [21]) for creating agents that adapt to a user's culture [1]. The agents primarily aim to increase senses of likeability, trust, and empathy with users by emulating the familiar social customs and behaviors based on models of cultural dimensions [28, 29, 47]. The models may be enough for these purposes, but is not enough to tell us how people would respond to specific feedback and how to best tailor the feedback to increase its effectiveness. For instance, some cultures may need more direct and straightforward feedback to achieve a positive result, while others may be best suited to a more indirect, motivational style. In our study, we saw both differences in attitudes to feedback and general differences in task load that would necessitate different approaches to giving criticism.

We encourage designers to adopt a user-centered approach to agent feedback: thus, the feedback's effect with regards to different cultures can be enhanced. Generalized models are good for just that—identifying common attitudes that govern general behaviors. However, for more task-oriented and complex scenarios (such as teaching or motivation), we would suggest first testing with users to understand how their cultural backgrounds affect their response to feedback. Based on these observations, an agent's feedback can be adjusted to ensure a positive outcome with users. These approaches, when used as a supplement to generalized models, have also been shown to be effective in creating appropriate nonverbal behavior in agents [15, 37]. However, more research is needed to understand exactly how to tailor feedback for different cultures such that they help motivate users towards accomplishing goals.

7 LIMITATIONS AND FUTURE WORK

We looked at three different levels of agent appearance (Human, Mouse, and Robot). However, we found that a limiting factor in our study involved the anthropomorphic similarities among each agent's appearance (i.e., sharing similar expressions and features, differing in appearance of head and hands). Though the similarity among appearance masked possible perception differences, we show that agents with consistent human-like anatomical features do not differ in perception. Further evaluations should study agents that are less anthropomorphic and more true to their natural form.

As stated earlier, our study was also split into two interaction styles (interface button or speech). Though we did not report on these results in this paper for brevity, we do plan to run a comprehensive evaluation in future work. The findings had no effect on the results presented in this study and we believe it would be best presented independently from the work done in this paper.

8 CONCLUSION

In this paper, we presented results from a study investigating the effect of agent appearance and cultural background on users' reactions and attitudes towards agent feedback. We hypothesized that feedback would evoke stronger reactions for more human-like agents and that user cultural background would affect perceptions of the agent's feedback. Surprisingly, we showed that agent appearance did not significantly affect how users respond to feedback. Additionally, cultural background affected both perceptions of task load as well as reactions to criticism from the agent. Our findings show that understanding user responses to feedback is more complex than what generalized cultural models [20, 21] can predict. We hope our work helps inform the design of motivational agents that are both culturally acceptable and more effective.

REFERENCES

[1] Mashael Al-Saleh and Daniela M Romano. 2015. Culturally Appropriate Behavior in Virtual Agents. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference.* AAAI Press, 69–74.

[2] Ruth Aylett, Natalie Vannini, Elisabeth Andre, Ana Paiva, Sibylle Enz, and Lynne Hall. 2009. But that was in another country: agents and intercultural empathy. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '09).* International Foundation for Autonomous Agents and Multiagent Systems, 329–336.

[3] Jeremy N Bailenson, Eyal Aharoni, Andrew C Beall, Rosanna E Guadagno, Aleksandar Dimov, and Jim Blascovich. 2004. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on PRESENCE.* 1864–1105.

[4] Jeremy N. Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. 2005. The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments. *Presence: Teleoperators and Virtual Environments* 14, 4 (Aug. 2005), 379–393. https://doi.org/10.1162/105474605774785235

[5] Amy L. Baylor. 2009. Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (Dec. 2009), 3559–3565. https://doi.org/10.1098/rstb.2009.0148

[6] Amy L Baylor. 2011. The design of motivational agents and avatars. *Educational Technology Research and Development* 59, 2 (2011), 291–300.

[7] Amy L. Baylor and Soyoung Kim. 2008. The Effects of Agent Nonverbal Communication on Procedural and Attitudinal Learning Outcomes. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Helmut Prendinger, James Lester, and MitsuruEditors Ishizuka (Eds.). Springer Berlin Heidelberg, 208–214.

[8] Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp. 2012. A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Yukiko Nakano, Michael Neff, Ana Paiva, and MarilynEditors Walker (Eds.). Springer Berlin Heidelberg, 126–138.

[9] Cynthia L. Breazeal and Rodney Brooks. 2000. *Sociable Machines: Expressive Social Exchange between Humans and Robots*. Ph.D. Dissertation. USA. AAI0801833.

[10] Justine Cassell. 2000. *Embodied Conversational Agents*. MIT Press, 1–27. http://dl.acm.org/citation.cfm?id=371552.371554

[11] Kerstin Dautenhahn and Iain Werry. 2004. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition* 12, 1 (Jan. 2004), 1–35. https://doi.org/10.1075/pc.12.1.03dau

[12] Nick Degens, Gert Jan Hofstede, John Mc Breen, Adrie Beulens, Samuel Mascarenhas, Nuno Ferreira, Ana Paiva, and Frank Dignum. 2014. *Creating a World for Socio-Cultural Agents*. Springer International Publishing, 27–43. https://doi.org/10.1007/978-3-319-12973-0_2

[13] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, and et al. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068. http://dl.acm.org/citation.cfm?id=2617388.2617415

[14] Birgit Endrass, Elisabeth André, Matthias Rehm, and Yukiko Nakano. 2013. Investigating culture-related aspects of behavior for virtual characters. *Autonomous Agents and Multi-Agent Systems* 27, 2 (Sept. 2013), 277–304. https://doi.org/10.1007/s10458-012-9218-5

[15] Birgit Endrass, Ionut Damian, Peter Huber, Matthias Rehm, and Elisabeth André. 2010. Generating culture-specific gestures for virtual agent dialogs. In *International Conference on Intelligent Virtual Agents*. Springer, 329–335.

[16] Arjan Geven, Johann Schrammel, and Manfred Tscheligi. 2006. Interacting with Embodied Agents That Can See: How Vision-Enabled Agents Can Assist in Spatial Tasks. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/1182475.1182490

[17] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[18] Adam T. Hirsh, Steven Z. George, and Michael E. Robinson. 2009. Pain assessment and treatment disparities: A virtual human technology investigation. *Pain* 143, 1 (May 2009), 106–113. https://doi.org/10.1016/j.pain.2009.02.005

[19] Geert Hofstede. 1980. *Culture's consequences: International differences in work-related values*. Vol. 5. sage.

[20] Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture* 2, 1 (2011), 8. https://doi.org/10.9707/2307-0919.1014

[21] G Hofstede, GJ Hofstede, and Michael Minkov. 2010. Cultures and organizations: software of the mind.

[22] Katherine Isbister, Hideyuki Nakanishi, Toru Ishida, and Cliff Nass. 2000. Helper Agent: Designing an Assistant for Human-human Interaction in a Virtual Meeting Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. ACM, 57–64. https://doi.org/10.1145/332040.332407

[23] Takuto Ishioh and Tomoko Koda. 2016. Cross-cultural Study of Perception and Acceptance of Japanese Self-adaptors. In *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI '16)*. Association for Computing Machinery, 71–74. https://doi.org/10.1145/2974804.2980491

[24] Toshikazu Kanaoka and Bilge Mutlu. 2015. Designing a Motivational Agent for Behavior Change in Physical Activity. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 1445–1450. https://doi.org/10.1145/2702613.2732924

[25] Felix Kistler, Birgit Endrass, Ionut Damian, Chi Tai Dang, and Elisabeth André. 2012. Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces* 6, 1 (July 2012), 39–47. https://doi.org/10.1007/s12193-011-0087-z

[26] T. Koda and P. Maes. 1996. Agents with faces: the effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. 189–194. https://doi.org/10.1109/ROMAN.1996.568812

[27] Kristen E Link, Roger J Kreuz, Arthur C Graesser, Tutoring Research Group, et al. 2001. Factors that influence the perception of feedback delivered by a pedagogical agent. *International Journal of Speech Technology* 4, 2 (2001), 145–153.

[28] Samuel Mascarenhas, Nick Degens, Ana Paiva, Rui Prada, Gert Jan Hofstede, Adrie Beulens, and Ruth Aylett. 2016. Modeling culture in intelligent virtual agents. *Autonomous Agents and Multi-Agent Systems* 30, 5 (2016), 931–962.

[29] Samuel Mascarenhas, João Dias, Nuno Afonso, Sibylle Enz, and Ana Paiva. 2009. Using rituals to express cultural differences in synthetic characters. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '09)*. International Foundation for Autonomous Agents and Multiagent Systems, 305–312.

[30] Punya Mishra. 2006. Affective feedback from computers and its effect on perceived ability and affect: A test of the computers as social actor hypothesis. *Journal of Educational Multimedia and Hypermedia* 15, 1 (2006), 107–131.

[31] Jonathan Mumm and Bilge Mutlu. 2011. Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior* 27, 5 (2011), 1643–1650.

[32] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, 72–78. https://doi.org/10.1145/191666.191703

[33] T. Noma, L. Zhao, and N. I. Badler. 2000. Design of a virtual human presenter. *IEEE Computer Graphics and Applications* 20, 4 (July 2000), 79–85. https://doi.org/10.1109/38.851755

[34] S. Parise, S. Kiesler, L. Sproull, and K. Waters. 1999. Cooperating with life-like interface agents. *Computers in Human Behavior* 15, 2 (1999), 123–142. https://doi.org/10.1016/S0747-5632(98)00035-1

[35] Isabella Poggi. 2004. Transcultural believability in embodied agents: a matter of consistent adaptation. *Agent Culture: Human-Agent Interaction in a Multicultural World* (2004), 75.

[36] Matthias Rehm and Elisabeth André. 2008. From Annotated Multimodal Corpora to Simulated Human-Like Behaviors. In *Modeling Communication with Robots and Virtual Humans (Lecture Notes in Computer Science)*, Ipke Wachsmuth and GüntherEditors Knoblich (Eds.). Springer Berlin Heidelberg, 1–17.

[37] Matthias Rehm, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner, Yukiko Nakano, Toyoaki Nishida, and Hung-Hsuan Huang. 2007. The CUBE-G approach – Coaching culture-specific nonverbal behavior by virtual agents. In *Organizing and learning through gaming and simulation: proceedings of Isaga 2007*, Igor Mayer (Ed.).

[38] Matthias Rehm, Nikolaus Bee, Birgit Endrass, Michael Wissner, and Elisabeth André. 2007. Too close for comfort? adapting to the user's cultural background. In *Proceedings of the international workshop on Human-centered multimedia (HCM '07)*. Association for Computing Machinery, 85–94. https://doi.org/10.1145/1290128.1290142

[39] Jeff Rickel and W. Lewis Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence* 13, 4–5 (May 1999), 343–382. https://doi.org/10.1080/088395199117315

[40] Jong-Eun Roselyn Lee, Clifford Nass, Scott Brenner Brave, Yasunori Morishima, Hiroshi Nakajima, and Ryota Yamada. 2007. The case for caring colearners: The effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication* 57, 2 (2007), 183–204.

[41] Akikazu Takeuchi and Taketo Naito. 1995. Situated Facial Displays: Towards Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., 450–455. https://doi.org/10.1145/223904.223965

[42] Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242.

[43] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, 1–12. https://doi.org/10.1145/3290605.3300511

[44] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300511

[45] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[46] Sangseok You, Jiaqi Nie, Kiseul Suh, and S. Shyam Sundar. 2011. When the Robot Criticizes You...: Self-Serving Bias in Human-Robot Interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 295–296. https://doi.org/10.1145/1957656.1957778

[47] Zhe Zhang, Ha Trinh, Qiong Chen, and Timothy Bickmore. 2015. Adapting a geriatrics health counseling virtual agent for the chinese culture. In *International Conference on Intelligent Virtual Agents*. Springer, 275–278.