Effects of Virtual Human Appearance Fidelity on Emotion Contagion in Affective Inter-Personal Simulations

Matias Volonte, Sabarish V. Babu *Member, IEEE*, Himanshu Chaturvedi, Nathan Newsome, Elham Ebrahimi, Tania Roy, Shaundra B. Daily and Tracy Fasolino

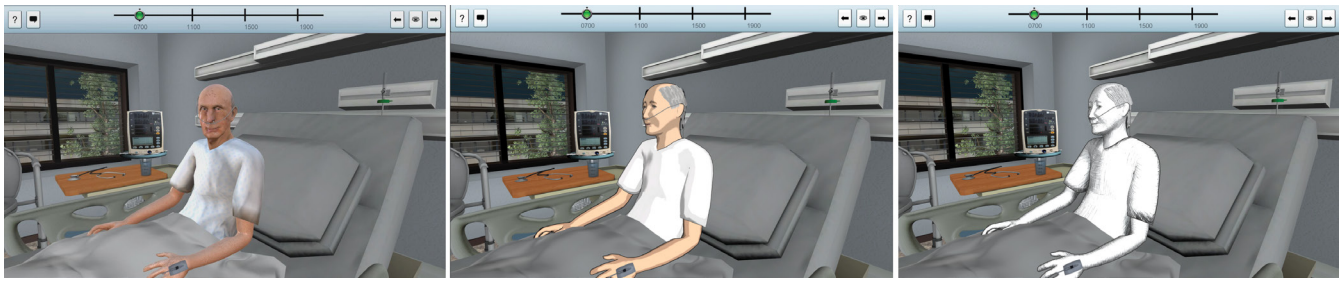

Fig. 1. Screenshots depicting the realistic (left) versus stylized cartoon like (center) and sketch like (right) representation of a virtual humans.

Abstract—Realistic versus stylized depictions of virtual humans in simulated inter-personal situations and their ability to elicit emotional responses in users has been an open question for artists and researchers alike. We empirically evaluated the effects of near visually realistic vs. non-realistic stylized appearance of virtual humans on the emotional response of participants in a medical virtual reality system that was designed to educate users in recognizing the signs and symptoms of patient deterioration. In a between-subjects experiment protocol, participants interacted with one of three different appearances of a virtual patient, namely visually realistic, cartoon-shaded and charcoal-sketch like conditions in a mixed reality simulation. Emotional impact were measured via a combination of quantitative objective measures were gathered using skin Electrodermal Activity (EDA) sensors, and quantitative subjective measures such as the Differential Emotion Survey (DES IV), Positive and Negative Affect Schedule (PANAS), and Social Presence questionnaire. The emotional states of the participants were analyzed across four distinct time steps during which the medical condition of the virtual patient deteriorated (an emotionally stressful interaction), and were contrasted to a baseline affective state. Objective EDA results showed that in all three conditions, male participants exhibited greater levels of arousal as compared to female participants. We found that negative affect levels were significantly lower in the visually realistic condition, as compared to the stylized appearance conditions. Furthermore, in emotional dimensions of interest-excitement, surprise, anger, fear and guilt participants in all conditions responded similarly. However, in social emotional constructs of shyness, presence, perceived personality, and enjoyment-joy, we found that participants responded differently in the visually realistic condition as compared to the cartoon and sketch conditions. Our study suggests that virtual human appearance can affect not only critical emotional reactions in affective inter-personal training scenarios, but also users' perceptions of personality and social characteristic of the virtual interlocutors.

Index Terms—Virtual/Digital Characters, Visual Fidelity, Emotion Contagion, Psychology, User Studies

1 Introduction

We have recently witnessed a rise in the development of human-like virtual agents, and a strong push to enhance their appearance and believability. Virtual humans are being deployed to simulate social interactions with humans. Medical educators, in particular, have become increasingly interested in using virtual humans for training systems that deal with interpersonal interactions that may be dangerous or difficult to reproduce in reality. Virtual humans have been used for many situations including: routine patient surveillance [10], negotiation simulation [7], interview emulation [8], interpersonal and leadership skills training [3], and uncomfortable situations such as prostate exams [29].

The visual fidelity of virtual humans is important to consider in both research and entertainment applications. Significant amount of time can be invested in 3D modeling, animation, and system design to create realistic human-like digital characters. Moreover, advanced computational resources dedicated to real-time rendering, shading and animation may be needed to render high fidelity virtual characters. Stylized, or "cartoon", rendering of virtual humans has been widely used, as it takes less time and resources to create. However, it has been unclear to what extent these rendering styles influence empathetic response in simulated interpersonal experiences.

It has been shown that human-agent interactions can evoke similar social responses to that of human-human interactions [29, 23]. Generating an emotional response from participants is important when working with learning systems, as it has a strong influence on memory retention [11]. If a system is able to elicit emotions from a participant, they may be more engaged and more likely to associate emotions with task interpretation. Research has shown that when the virtual character's responses were empathetic towards the participant's emotions,

- *Matias Volonte is with the School of Computing at Clemson University. E-mail: mvolont@clemson.edu.*
- *Sabarish V. Babu is with the School of Computing at Clemson University. E-mail: sbabu@clemson.edu.*
- *Himanshu Chaturvedi is with the School of Computing at Clemson University. E-mail: hchatur@clemson.edu.*
- *Nathan Newsome is with the School of Computing at Clemson University. E-mail: nathann@clemson.edu.*
- *Elham Ebrahimi is with the School of Computing at Clemson University. E-mail: eebrahi@clemson.edu.*
- *Tania Roy is with the School of Computing at Clemson University. E-mail: taniar@clemson.edu.*
- *Shaundra B. Daily is with the Department of Computer, Information Science and Engineering at the University of Florida. E-mail: shanib@ufl.edu.*
- *Tracy Fasolino is with the School of Nursing at Clemson University. E-mail: tfasoli@clemson.edu.*

Manuscript received 21 Sept. 2015; accepted 10 Jan. 2016. Date of publication 20 Jan. 2016; date of current version 19 Mar. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2518158

there was a significant decrease in stress and a positive effect on the participants' perception of the task [25]. These emotional responses can be used to measure co-presence in virtual environments to see if participants are forming a bond with the character. When working with medical training systems, it is especially important to accurately mimic human behaviors and characteristics in order to elicit empathetic responses from participants. Research shows a significant impact of the elicited emotional responses on participants' learning outcomes for a task with virtual humans [29]. Furthermore, evidence suggests that the emotional state of medical practitioners can have an impact on their professional abilities [30].

Avoiding the Uncanny Valley phenomenon has been a challenge in the computer animation and robotics fields for some time [21, 15]. The theory suggests that empathetic response increases with the human-likeness of the virtual human to a certain point and then drops to a point of revulsion before increasing again. The entertainment industry has provided a great example for this phenomenon. Levy [17] discusses how audiences found the highly realistic human characters in Polar Express to be creepy and off-putting, while most found the stylized cartoon characters in Pixar's The Incredibles more appealing. While there have been significant strides in understanding the phenomenon, it is our intention to examine the impact of visual fidelity of virtual humans on this phenomenon when measured using quantitative and qualitative methods examining the emotional responses of users.

For a previous study, we developed a medical virtual reality training system to help nurses identify the signs of rapid patient deterioration [38]. While developing the system, we worked closely with medical experts to create accurate virtual humans based on actual patient data. In this work, we used this system to study how different rendering techniques (realistic and non-realistic) emotionally impacted participants. Our empirical evaluation adds to the current understanding of the effects of visual fidelity of virtual humans on human emotional responses.

The current contribution extends our previous work [4] that presented the effects of visual fidelity of virtual humans on positive and negative affect. In this research contribution, based on feedback from the graphics and VR scientific community, we have drastically improved the virtual human appearance in the realistic condition by using advanced modeling, shading and rendering techniques to enhance the human-like resemblance of the virtual patient. In a comparative empirical evaluation, we have investigated the effects of three visual appearance conditions (cartoon, sketch and visually realistic) of the virtual patient on not only the positive and negative affect of the participants, but also to include other quantitative measures such as objective skin electrodermal activity, social presence, differential emotion survey scores and performance variables.

2 Related Work

Presence of emotional effects in human-virtual agent interactions has already been exhibited in previous studies [23, 29]. De Melo et al. [9], observed that people attempted methods of emotional negotiation with virtual humans. It has also been established that users do not disregard their social standards and stereotypes when interacting with virtual agents [22]. Researchers have studied whether synthetic emotions expressed by virtual humans elicit emotion in a human conversation partner [27]. They found that negative compared to positive synthetic emotions expressed by a virtual human can elicit a more negative emotional state in a human conversation partner. Effects of virtual agents on participants while performing tasks were examined [39] and it was found that participants were socially inhibited while performing complex tasks, and social facilitation did not occur while doing simple tasks in the presence of virtual humans. Research has also focused on enhancing the expressiveness of virtual characters so that the users respond to interpersonal simulations in the intended manner. Various models of individual virtual character's emotion, models of interpersonal behavior and methods of generating expression for creating virtual characters with higher fidelity have been discussed in [35]. Attempts have been made to understand the characteristics of virtual agents which have the greatest effect on users' emotions. Tsai et al. [32] showed that a still image of a virtual agent could invoke empathy in the user when comparing a happy expression with that of a neutral one. In a previous study, we examined the effect of the presence of animation on emotion contagion in a medical interpersonal simulation with virtual patients [38]. Our research showed that the presence of the animation had a significant effect on users' emotional responses, especially negative emotions corresponding to the simulated deterioration of the virtual patient, and it also enhanced the levels of social presence in user's interactions with the virtual patients.

Quite a few studies have been done to understand the relation between various rendering techniques of virtual humans and the emotional response of the participants. Mandryk et al. [19] studied subjective emotional responses to multiple non photorealistic rendering (NPR) approaches of virtual humans in digital videos, and found that NPR algorithms dampened participants' emotional responses in terms of arousal and valence. McDonnell et al. [20] carried out experiments to determine the effect of render styles on perception of personality of virtual humans. They found that the characters that appear highly abstract or highly realistic are considered equally appealing. Extremely unappealing characters appear even less appealing when moving than when still. Ellis et al. [12] demonstrated that the extent to which subjects ascribe emotions to VR faces is highly dependent on textures (photo-realistic vs. non-photo-realistic) applied to the face. They found no significant difference for the happy or surprised conditions and all the variance resulted from the sad and angry conditions, both of which favor the photo-realistic face.

MacDorman et al. [18] studied different qualities of rendered models of virtual characters namely photo-realistic, bronze and line drawings in order to examine the perceived eeriness and human likeness when participants were exposed to vignettes of these virtual characters. One of the interested finding of this research suggests that increasing the model fidelity of the virtual humans by enhancing the polygonal level of detail may decrease the perceived eeriness in the photo-realistic condition. Bailenson et al. [2] investigated how co-presence in an immersive virtual environment is influenced by variations in how much an embodied agent resembles a human being in appearance and behavior. The results of their research seems to indicate that co-presence was lowest when there was a large mismatch between the appearance and behavioral realism of an embodied agent. Ring et al. [28] studied the relationship between the virtual agent's rendering style (realistic or toon shaded) and the task domain (medical or social) on users' ratings of friendliness, likability, caring, and appeal depending on the content of the dialogue. Overall, they found that the agent's appearance affected the viewers impressions, and they found that toon shaded characters were rated as being more likable, and caring when compared to realistic characters in social task contexts.

More recently, Wellerdiek et. al. [37] found that the body shape and pose of virtual characters can potentially affect the perception of physical strength and social power. They found that characters with a weaker looking body shape can be perceived as more powerful when presented in a high-power pose, as compared to a stronger looking body shape. In our experiment design, when modeling the visually realistic appearance of the virtual human, we strived to match the body shape and actions of the realistic appearing virtual human to exactly that of the stylized appearance conditions so that these unforeseen effects, such as perceived strength and power do not confound the measured emotional reactions of the participants. Of relevance is the research of Zibrek and McDonnell [40] in which the authors found that the rendering style of the virtual character have the potential to affect the perceived personality of the character. They found that a visually realistic but ill-looking rendering style evoked a less desirable (less agreeable) personality than the cartoon style presentation of a virtual human. These results have implications to how participants may perceive ill-looking yet visually realistic virtual humans in clinical deterioration scenarios in interactions with medical virtual reality systems like our RRTS. Hyde et. al. [16] investigated the effects of eye gaze and visual realism of virtual humans on the perception of personality. They found that if the eye gaze and blinking behavior of a virtual character were accurately modeled then participants perceive the char-

acters personality to be robust and consistent, despite the differences in visual realism of the character models. In our experiment, we strived to maintain consistency in the animation quality of facial expressions and gestures, such that our experiment setup can allow us to examine the effects of visual appearance fidelity alone on the emotional reactions of participants in interactions with a virtual human.

Although emotion contagion and virtual human perception studies have been done in the past, most of them have used virtual characters in the form of still images or in short video clips. To our knowledge, this is the first research to study the effect of realistic vs. stylized appearance on users' emotion contagion in an interactive virtual environment (IVE).

3 Experiment Simulation and Setup

The Rapid Response Training system (RRTS) [10] was designed to train nurses in recognizing the signs and symptoms of rapid deterioration in patients. The RRTS simulates the daily duties of a nurse including visiting multiple patients typically four times a day, gathering the vital signs of the patients and reporting them in an Electronic Health Record System (EHR). To provide this experience our system is setup in a dual screen configuration. A large screen display is used to show the virtual hospital environment which includes the patient and the various instruments present in a general ward in a life-size view. The second display is used to present the Electronic Health Record [EHR] that would be used to report the quantitative and qualitative vital signs gathered while interacting with the virtual patient.

Virtual humans used in the RRTS have been modeled after real life patients who have undergone rapid deterioration, over the course of a nurse's shift. The signs and symptoms of deterioration have been carefully modeled and animated in our virtual patients with the help of medical experts. Several modes of interaction with the patient are provided. The participants can ask a set of pre-defined medically relevant questions (via a dialogue box) or use one of the instruments in the patient's environment to measure his vital signs, and record the observations in the simulated EHR system. Detailed information regarding the RRTS design, virtual human and instrument interaction and implementation can be found in [10, 38]. In this research, the RRTS served as a rich experiment platform for empirical examination of how factors associated with the appearance of the virtual human can affect the users' emotional state and responses (see Figure 2).

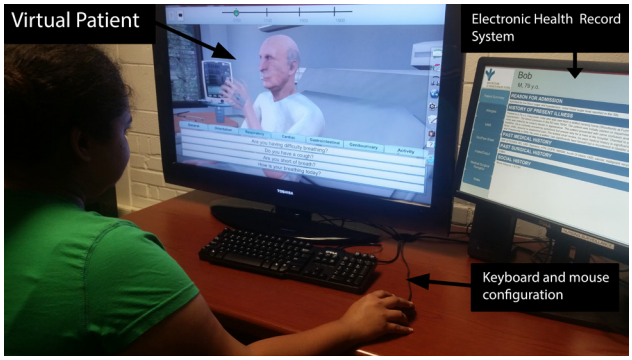

Fig. 2. Screenshot shows a participant interacting with the virtual patient Bob in the RRTS, and recording his vitals in the EHR screen.

The experiment featured three conditions, one realistic and two non-realistic (Cartoon and Sketch). When considering the rendering style of our non-photorealistic (NPR) conditions, we used two very distinct techniques. After assessing many possibilities, we modeled one NPR condition to be cartoon-like (cel shaded), and the other to appear hand drawn as a sketch. We aimed to provide a consistent user experience among experiment conditions with the RRTS, and avoid any possible confounds due to differences in virtual environment and instrument appearance. Therefore, the only modeled difference in the experiment conditions were the rendering techniques applied to the virtual patient; all virtual human animations and behaviors, as well as the look and feel of the virtual environment and instrument interactions in RRTS remained constant (see Figure 1).

In the realistic condition, the skin textures of the virtual patient include details such as wrinkles and blemishes to increase realism. A normal map created in Zbrush was used to depict skin wrinkles. Moreover, the face texture resolution in the realistic condition was 2908 height x 2908 width while it was 1024 height x 1024 width in the non-realistic conditions. Furthermore, high definition range image-based lighting was implemented for the realistic condition. Finally, the mesh used for the characters head was remodeled and the polygon count was increased. In the realistic condition, characters head has a total of 30,227 polygons vs 2,728 polygons in the charcoal sketch and toon conditions.The Cartoon, or cel shaded, condition used Unity3D's Basic Outline Toon shader that gives the character a uniform outline and simplified two color shading. Only the skin and eyeball textures were kept in detail in order to provide an outline of the eyes, eyebrows, and pupils. Specular highlights provided visual clues about the material type (skin, clothing etc) of the model. Unlike standard shaders, there are clear borders for shadows on cel shaded objects. The Sketch condition was implemented using a custom shader to give the character a hand-drawn charcoal sketch like appearance. A drawing technique called hatching was used, which refers to closely spaced parallel strokes that follow the curvature of a surface to define volume and materiality. No color is used in this condition, instead the density of the hatch marks defines tone, where less dense strokes denotes a lighter tone and denser strokes a darker tone. Cross-hatching is also applied to create darker tones by layering strokes at different angles. We created a texture based shader that used a pre-rendered sequence of mip-mapped hatch stroke images which correspond to different tones. A multi-texturing algorithm was implemented to calculate the lighting tones for each face, and then blend the appropriate hatching images to render the proper shading to the model of the virtual patient. Each texture in the sequence contains the hatch marks from all the previous images. This nesting property helped to provide seamless blending between tones.

The RRTS was deployed in a dual screen system that used a 60" TV as the main screen for the simulation to display the interactive virtual patient (Bob) and a 21" LCD Monitor for the EHR system (as shown in Figure 2). The participant interacted with the system using a standard mouse and keyboard. They wore a sensor (Qsensor from Affectiva Inc.) that measured the skin electro-dermal activity on the wrist of their non-dominant hand. A separate laptop was provided in between time-steps for gathering responses to the surveys.

4 Experiment Procedure

4.1 Research Question and Expected Outcomes

A good deal of time and forethought went into creating realistic human-like appearance for our virtual characters in the RRTS. We found ourselves wondering what effect all this extra effort had on the user. We narrowed down three essential questions that are currently unanswered in the research literature:

- To what extent does the visual appearance fidelity of virtual humans affect the emotional reactions of users?
- To what extent does a photo-realistic (PR) human-like appearance of a virtual human versus two levels of a non-photo-realistic (NPR) appearance of a virtual human, namely a cartoon like rendering and a charcoal sketch like rendering, affect the emotional responses of users?
- To what extent does amplifying the emotional behaviors of the virtual human impact the users' affective responses in each of the rendering conditions?

Our initial hypothesis and expected outcomes were as follows: We hypothesize that *participants in the visually realistic condition might exhibit greater levels of arousal and negative emotional response as*

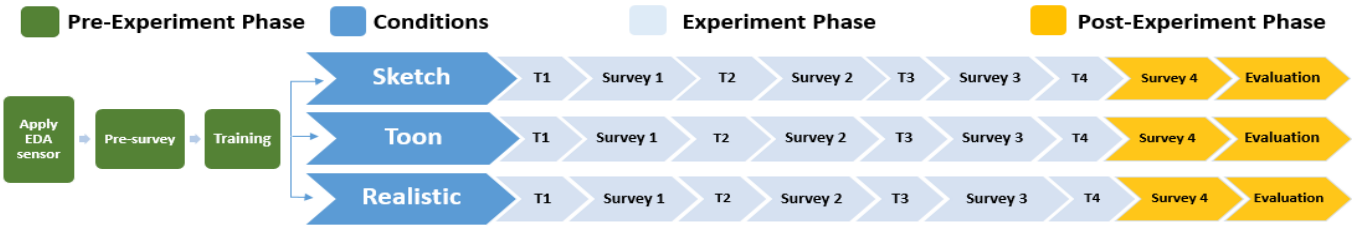

Fig. 3. The time-line of the experiment from left to right.

compared to participants in the NPR Cartoon and Sketch shaded conditions, corresponding to the gradual medical deterioration and distress of the virtual patient Bob in the RRTS. However, we do not expect to find any differences in participants' positive affect, sense of presence or performance related variables in the visually realistic versus stylized appearance conditions of Bob.

4.2 Study Design

In order to empirically examine each of these questions, we developed a 3x4 experiment. We had three between subjects conditions: Realistic rendering, Cartoon rendering, and Sketch rendering. There were four within subject conditions. These were the distinct time-steps that had increasing levels of negative stimuli. Each participant was exposed to the same virtual agent, Bob, in each of the four time-steps. The declining health condition of Bob was expressed through his vital signs, verbal feedback, and non-verbal behaviors. The animated behaviors of the agent changes between time stamps to reflect this decline in the medical condition of Bob.

We recruited both male and female participants, between the ages of 18 and 35, who had a basic knowledge of the medical terminology present in the simulation. We ran a total of 62 participants (20 in Realistic, 22 in Cartoon, and 20 in Sketch rendering conditions), who were recruited from the College of Engineering and Science at Clemson University. We had a near equal distribution of gender in our participant pool that included 36 males and 26 females (Realistic: M 11, F 9, Cartoon: M 10, F 12, Sketch: M 15, F 5).

4.3 Methodology

Participants first listened to a brief explanation regarding the design and objectives of the RRTS system. After consent was obtained, we placed an EDA Q Sensor on the wrist of the participant's non-dominant hand. The participant was then asked to fill out a series of questionnaires regarding his or her background and current disposition. The participant was trained on the system using a virtual human other than the one that would be used throughout the experiment, but in the same visual rendering condition (Realistic or Cartoon or Sketch) exhibiting a stable medical condition (time-step one). The experimenter introduced the participant to every interaction in the virtual environment during the training phase before initiating the study. Then the participant was introduced to the first time-step and asked to interact with Bob by asking as many questions as possible, use as many virtual instruments as necessary to medically assess Bob's condition, and record the observations in the EHR system. At the end of each time-step, the participant filled out a Differential Emotions Survey (DES) and a Positive and Negative Affect schedule (PANAS) questionnaire to reflect his or her current emotional state. The participant was also given a short quiz on the condition of the patient after every time-step as an incentive to increase interactions with Bob. Finally, at the end the fourth time-step, after they were administered the PANAS and DES surveys, they completed the social presence survey, and were debriefed and thanked for their time. Figure 3 illustrates the various stages of the experiment protocol.

4.4 Measures

The independent variables are visual appearance conditions of the virtual human, and the intensity of the negative emotions of the virtual agent from one time-step to the next of the simulated deterioration scenario of the virtual patient. There were a number of dependent variables used to obtain the emotional reactions of the user.

4.4.1 Quantitative Measures

Any physiological arousal, both external and internal, affects the sympathetic nervous system, which controls the sweat producing eccrine glands and makes the skin a better conductor of electricity [5]. Within the sympathetic and parasympathetic division of the autonomic nervous system, balanced activity helps to regulate physiological states of arousal. The parasympathetic nervous system (PNS) and sympathetic nervous system (SNS) are often compared to the brakes and gas of a car respectively; helping "rest and digest" versus "fight and flight" activities [6]. When SNS activity increases, the fibers around the eccrine sweat glands regulate the production of sweat. This measurable quantity of electrical conductance is known as electrodermal activity (EDA), and can be measured as skin conductance using wrist worn sensors [24]. In this case, the wireless Q sensor, manufactured by Affectiva, was placed on the ventral side of the participant's wrist of the non-dominant hand. To mark the different time steps in this study (Figure 3), the timestamp button on the sensor was used to indicate the start and the stop time. Data collected from the participant, was transferred from the sensor to a computer using a USB cable. To account for between subject variance and baseline fluctuations the EDA data was normalized, using the formula below, as described by Healey et al [14]. SCR refers to Skin Conductance Response.

$$\frac{SCR - MEAN(SCR)}{MAX(SCR) - MIN(SCR)} \tag{1}$$

For every participant, the mean and standard deviation values were calculated from the normalized EDA data set. The window for each mean and standard deviation were the timestamps mentioned above in section 4.3 (demographic questionnaire, training phase, and timestamp1 to timestamp 4). The same process was followed for all the three experimental conditions (Toon, Sketch and Realistic).

The Positive and Negative Affect Schedule was used to evaluate an overall valence of emotion [36]. This is a 20 item survey, with each item being scored on a Likert scale from 1 to 5. The score reflects how strongly the participant felt an emotion at a given time, 1 being not at all and 5 being extremely. The results were grouped by positive and negative items.

The Differential Emotions Scale (DES) questionnaire in our experiment was based on the DES IV with a modification to reduce the item count to 30 [33]. There are 10 categories scored in the DES (3 items per category). The participant used a 0-9 Nominal style scale to express how strongly they felt each item. Again, 0 being Never and 9 being Extreme. The results were grouped into categories, and each category received its own score. The ten categories used in our questionnaire are as follows: Interest, Enjoyment, Surprise, Sadness, Anger, Contempt, Fear, Guilt, Shame and Shyness.

In addition to the repeated measures above, there was a final evaluation of the immersive virtual patient Bob. This survey contained 5 items based on the social presence survey [1]. These were scored on a 1-7 Nominal style scale, 1 being not at all and 7 being a great deal.

4.4.2 Qualitative Measures

In an attempt to better interpret the quantitative items, we used a number of open-ended discussion questions. These questions were used to assess their overall experience in the simulation. Two examples of these questions were: what did you like most about Bob? and how would you describe Bob's personality and appearance?

5 Results

When Analysis of Variance ANOVA analysis was conducted in our empirical evaluation, we first tested to ensure that the scores were normally distributed and Mauchley's test of sphericity has been satisfied. In cases where the test of sphericity was significant (assumption of homogeneity of variance was violated), Greenhouse-Geisser corrected degrees of freedom was used to assess the significance of the corresponding F values. Tukey HSD post-hoc was used to evaluate the effects of the between subjects variable of rendering style within each time-step as well as between levels of gender, and Bonferroni adjusted type 1 error (alpha) was used for within subjects post-hoc comparisons across time within scores in a rendering style. In pairwise parametric tests, Levene's test of equality of variance was conducted, and degrees of freedom were adjusted when variances were not assumed to be equal. Effect size measures are reported where possible.

5.1 Quantitative Results

5.1.1 Objective Measure: Skin Electrodermal Activity

In order to measure if the arousal levels of the participants differed by the rendering style of the virtual human or gender of the participant, and to examine the differences in arousal levels based on the deterioration scenario presented in the RRTS, we conducted a 3 x 2 x 6 mixed model repeated measures Analysis of Variance (ANOVA) on the mean EDA scores with sampling time across six different periods in the experiment. The different periods tested consisted of the pre-experiment demographics survey, training, and time-steps 1 to 4, as a within-subjects repeated measures variable. The conditions Visually Realistic (R), Sketch (S) and Cartoon (C) shading were the between-subjects variable. The ANOVA analysis did not reveal a main effect of time or condition, as well as time by condition and time by gender interactions were not significant. However, the three way sampling time by gender by condition interaction was significant $F(3.62, 92.21) = 3.20$, $p = 0.020$, $\eta^2 = 0.11$ (obs. power = 0.78), and the main effect of gender was significant $F(1, 51) = 12.90$, $p = 0.001$, $\eta^2 = 0.20$ (obs. power = 0.94).

In order to further examine the three way interaction effect overall between male and female participants, we conducted a block analysis of mean EDA scores across sampling times separately in males and females, and between mean EDA within sampling times between males and females. A one-way repeated measures ANOVA of EDA scores across sampling times in females was significant $F(1.82, 45.46) = 3.98$, $p = 0.029$, $\eta^2 = 0.14$ (obs. power = 0.66), whereas a similar analysis on mean EDA scores for males did not reveal a significant difference. Pairwise comparisons of mean EDA scores in females across sampling times using the Bonferroni adjustment method did not reveal a significant difference. Next, at each of the six sampling times, we conducted a pairwise comparison between male and female mean EDA scores that revealed that the mean EDA scores were significantly higher in males than females in the following time steps (See Figure 4): Training session males (M=0.40, SD=0.25) and females (M=0.22, SD=0.16), $t(51.28) = 3.16$, $p = 0.003$; Time-step 1 males (M=0.40), SD=0.26) and females (M=0.21, SD=0.15), $t(49.61) = 3.10$, $p = 0.003$; Time-step 2 males (M=0.39), SD=0.26) and females (M=0.19, SD=0.14), $t(49.77) = 3.74$, $p = 0.001$; Time-step 3 males (M=0.41), SD=0.26) and females (M=0.18, SD=0.13), $t(45.43) = 4.38$, $p < 0.001$; Time-step 4 males (M=0.41), SD=0.27) and females (M=0.18, SD=0.15), $t(47.88) = 4.06$, $p < 0.001$. Overall, mean EDA scores for males were higher than females in all the time steps of the interaction with the RRTS.

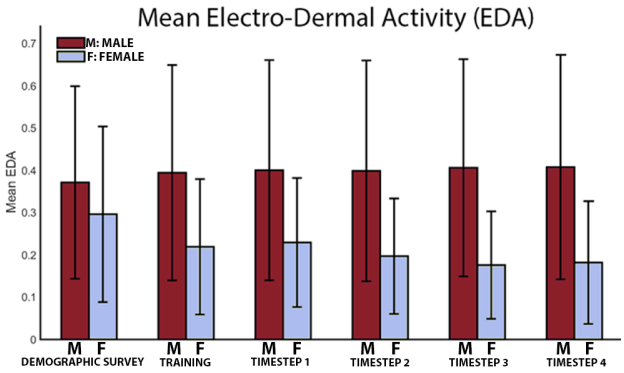

Fig. 4. Shows a graph of mean EDA scores between males and females across different time-steps of the participants interaction with the RRTS.

5.1.2 Subjective Measure: Positive and Negative Affect Schedule

In order to assess if there were any significant effects of condition or gender on positive and negative affect scores across the five sampling times, namely pre-experiment, and after time steps 1 through 4 corresponding to Bob's deterioration, we performed a 3 x 2 x 5 mixed model repeated measures ANOVA. The between subjects variable were condition (3 levels) and gender (2 levels), and the within subjects variable were the sampling times. The ANOVA analysis did not reveal any significant main effects or interaction effects for positive affect scores. With respect to mean negative affect scores, the ANOVA analysis revealed a significant main effect of condition $F(2, 56) = 3.31$, $p = 0.044$, $\eta^2 = 0.10$ (obs. power = 0.60), a significant main effect of time $F(1.73, 97.12) = 32.72$, $p < 0.001$, $\eta^2 = 0.37$ (obs. power = 1.0), and a significant time by condition interaction $F(3.47, 97.12) = 2.72$, $p = .041$, $\eta^2 = 0.09$ (obs. power = 0.69). Pairwise comparisons revealed that overall negative affect was the lowest in condition R (M=1.38, SD=0.10) as compared to C (M=1.73, SD=0.9) $p < 0.040$ (See Figure 6a).

In order to further examine the interaction effect, block analyses were conducted on mean negative affect scores within each virtual appearance conditions (R, S, C) examining differences across time-steps, and between different conditions within each time-step. In condition C, mean negative affect scores after time-step4 (M=2.71, SD=1.20) were significantly higher than after baseline (M=1.30, SD=0.29) $p < 0.001$, time-step1 (M=1.46, SD=0.50) $p = 0.002$, time-step2 (M=1.45, SD=0.57) $p = 0.001$, and time-step3 (M=1.73, SD=0.84) $p = 0.021$. In condition S, mean negative affect scores after time-step4 (M=2.45, SD=1.16) were significantly higher than after baseline (M=1.17, SD=0.15) $p = 0.001$, time-step1 (M=1.31, SD=0.40) $p = 0.007$, time-step2 (M=1.44, SD=0.51) $p = 0.019$, and time-step3 (M=1.49, SD=0.51) $p = 0.019$. Finally, in condition R, mean negative affect scores after time-step4 (M=1.74, SD=0.88) were significantly higher than after baseline measure (M=1.23, SD=0.32) $p = 0.026$. Comparisons of mean negative affect scores between the visual appearance conditions of Bob using Tukey-HSD analysis within each time-step of the interaction with the RRTS revealed that mean negative affect scores of participants in condition R (M=1.74, SD=0.88) were significantly lower as compared to participants in condition C (M=2.71, SD=1.20) $p = 0.017$ in time-step4. Overall, mean negative affect scores seemed to be lower in the visually realistic condition as compared to cartoon and sketch shaded condition at every time-step, and in all conditions negative affect seemed to gradually increase from baseline to time-step4 corresponding to the medical deterioration of the virtual patient.

5.1.3 Differential Emotions Survey

In order to examine the dimensions of emotion response of participants to the deterioration scenario of Bob at a more granular level, we administered the Differential Emotions Survey (DES) at the end of every time-step of interaction with Bob. We also wanted to examine if

the visual rendering conditions of the virtual human had any effect on the affective responses of participants, corresponding with the medical deterioration scenario, in each of the 10 dimensions of the DES survey. In each of ten granular dimensions of the DES questionnaire, we employed the following analyses. We first performed a 3 x 2 x 4 mixed model repeated measures ANOVA, with gender (2 levels) and visual appearance conditions (3 levels) as the between-subjects variables, and the scores on each of the DES dimensions after each time-step of interaction with Bob, as a within-subjects repeated measures variable. This tested for the main effect of condition, gender, DES dimension, gender by DES interaction, and the condition by DES interaction terms. Appropriate post-hocs and block analyses were conducted for follow-ups on main and interaction effects respectively (same as Section 5.1.2).

In the dimension ***interest-excitement***, the mixed model ANOVA analysis only revealed a significant main effect of sampling time of interest-excitement scores across the different time-steps, $F(2.21, 118.97) = 60.51$, $p < 0.001$, $\eta^2 = 0.52$ (obs. power = 1.0). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that overall interest-excitement scores of participants were significantly different from one another across each time-step of the interaction with the virtual patient Bob (See Figure 5a). Mean interest-excitement after time-step1 (M=4.30, SD=1.91) was the highest, as compared to after time-step2 (M=1.44, SD=1.60), after time-step3 (M=1.21, SD=1.43), and after time-step4 (M=2.64, SD=1.95).

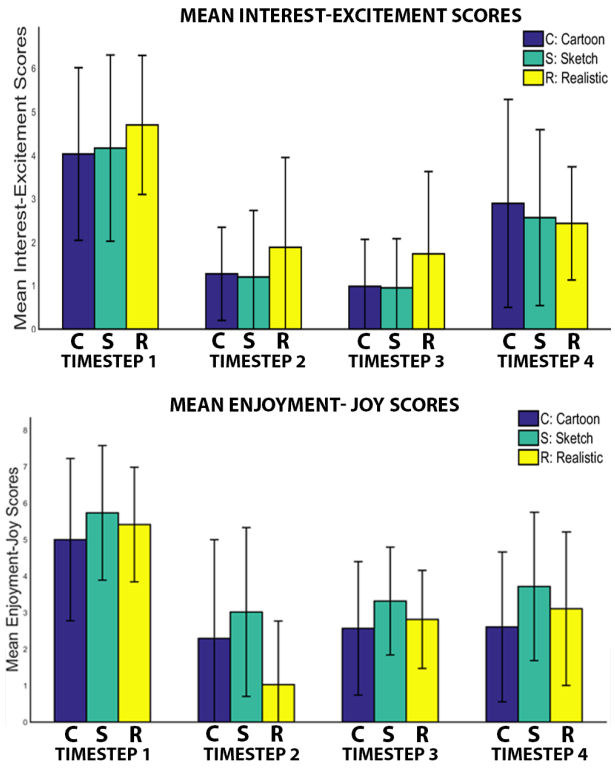

Fig. 5. Graph showing mean positive affect dimensions of DES of Interest-Excitement scores (a) and Enjoyment-Joy scores (b) for participants across time-steps by visual appearance conditions of the virtual patient.

In the dimension ***enjoyment-joy***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of enjoyment-joy scores across the different time-steps $F(2.06, 115.18) = 71.31$, $p < 0.001$, $\eta^2 = 0.56$ (obs. power = 1.0), and a time by condition interaction $F(4.11, 115.18) = 5.14$, $p = 0.001$, $\eta^2 = 0.15$ (obs. power = 0.96). In order to further examine the sampling time by visual appearance condition interaction on the mean enjoyment-joy scores, we conducted multiple block analyses (See Figure 5b). In condition C, pairwise post-hoc comparisons of mean enjoyment-joy scores conducted via Bonferroni analysis showed that the scores were highest after time-step1 (M=5.21, SD=2.12) and differed significantly from scores after time-step2 (M=2.93, SD=2.69) $p = 0.002$, time-step3 (M=2.92, SD=1.87) $p < 0.001$, and time-step4 (M=2.98, SD=2.01) $p < 0.001$. In condition S, pairwise comparisons showed that the mean enjoyment-joy score were highest after time-step1 (M=5.38, SD=1.88) and differed significantly from scores after time-step2 (M=2.32, SD=2.39) $p < 0.001$, time-step3 (M=2.98, SD=1.48) $p < 0.001$, and time-step4 (M=3.11, SD=2.29) $p = 0.001$. In condition R, pairwise comparisons showed that the mean enjoyment-joy scores were similarly highest after time-step1 (M=5.63, SD=1.52), and differed significantly from scores after time-step2 (M=0.70, SD=1.04) $p < 0.001$, time-step3 (M=2.81, SD=1.21) $p < 0.001$, and time-step4 (M=3.41, SD=2.01) $p < 0.001$. Time-step2 was the lowest and also significantly differed from time-steps 3 $p < 0.001$ and 4 $p < 0.001$. Mean enjoyment-joy scores were not significant different between conditions C, S, and R after time-step1 or time-step3 or time-step4. However, Tukey-HSD post-hoc analysis revealed that mean enjoyment-joy scores at time-step2 were significantly lower in the visually realistic condition (M=0.70, SD=1.04) than the cartoon condition (M=2.93, SD=2.69) $p = 0.004$, with the scores in the sketch condition in the middle (M=2.32, SD=2.39).

In the dimension ***surprise-startle***, the mixed model ANOVA analysis did not reveal any significant effects. In the dimension ***distress-anguish***, the mixed model ANOVA analysis only revealed a significant main effect of sampling time of distress-anguish scores across the different time-steps overall, $F(2.52, 142.14) = 6.96$, $p < 0.001$, $\eta^2 = 0.11$ (obs. power = 0.96). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean distress-anguish scores after time-step3 (M=1.98, SD=2.01) was the lowest, and significantly different as compared to after time-step1 (M=2.99, SD=1.92) $p = 0.006$, after time-step2 (M=3.19, SD=1.46) $p < 0.001$, and after time-step4 (M=2.90, SD=2.18) $p = 0.002$ corresponding to the medical deterioration of the virtual patient Bob in all visual appearance conditions (See Figure 6b).

In the dimension ***anger***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of anger scores across the different time-steps overall, $F(2.59, 145.16) = 11.28$, $p < 0.001$, $\eta^2 = 0.17$ (obs. power = 0.99). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean anger scores after time-step1 (M=0.92, SD=1.24) was the lowest, and significantly different as compared to after time-step2 (M=2.23, SD=2.22) $p < 0.001$, after time-step3 (M=2.46, SD=2.17) $p < 0.001$, and after time-step4 (M=1.99, SD=2.26) p = 0.001 corresponding to the medical deterioration of the virtual patient Bob in all visual appearance conditions (See Figure 6c).

In the dimension ***contempt-scorn***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of contempt-scorn scores across the different time-steps overall $F(1.91, 106.97) = 11.69$, $p < 0.001$, $\eta^2 = 0.17$ (obs. power = 0.99), and a sampling time by gender interaction $F(1.91, 106.97) = 3.78$, $p = 0.27$,$\eta^2 = 0.06$ (obs. power = 0.66). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean contempt-scorn scores after time-step4 (M=2.19, SD=1.75) was the highest, and significantly different as compared to after time-step2 (M= 0.92, SD=1.27) $p < 0.001$, and after time-step3 (M=1.03, SD=1.39) $p < 0.001$ corresponding to the medical deterioration of the virtual patient Bob in all visual appearance conditions. Pairwise comparisons using Tukey-HSD revealed that mean contempt-scorn scores in males (M=1.92, SD=2.00) were significantly higher than females (M=0.87, SD=1.64) at time-step1 $p = 0.032$ with respect to how they perceived Bob overall (See Figure 6d).

In the dimension ***fear-terror***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of fear-terror scores across the different time-steps overall, $F(2.05, 114.76) = 29.77$, $p < 0.001$, $\eta^2 = 0.35$ (obs. power = 1.0). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean fear-terror scores after time-step4 (M=3.41, SD=2.46) were the highest corresponding Bob's worst medical condition, and was significantly different as compared to after time-step1 (M=1.72, SD=1.36) $p < 0.001$, after time-

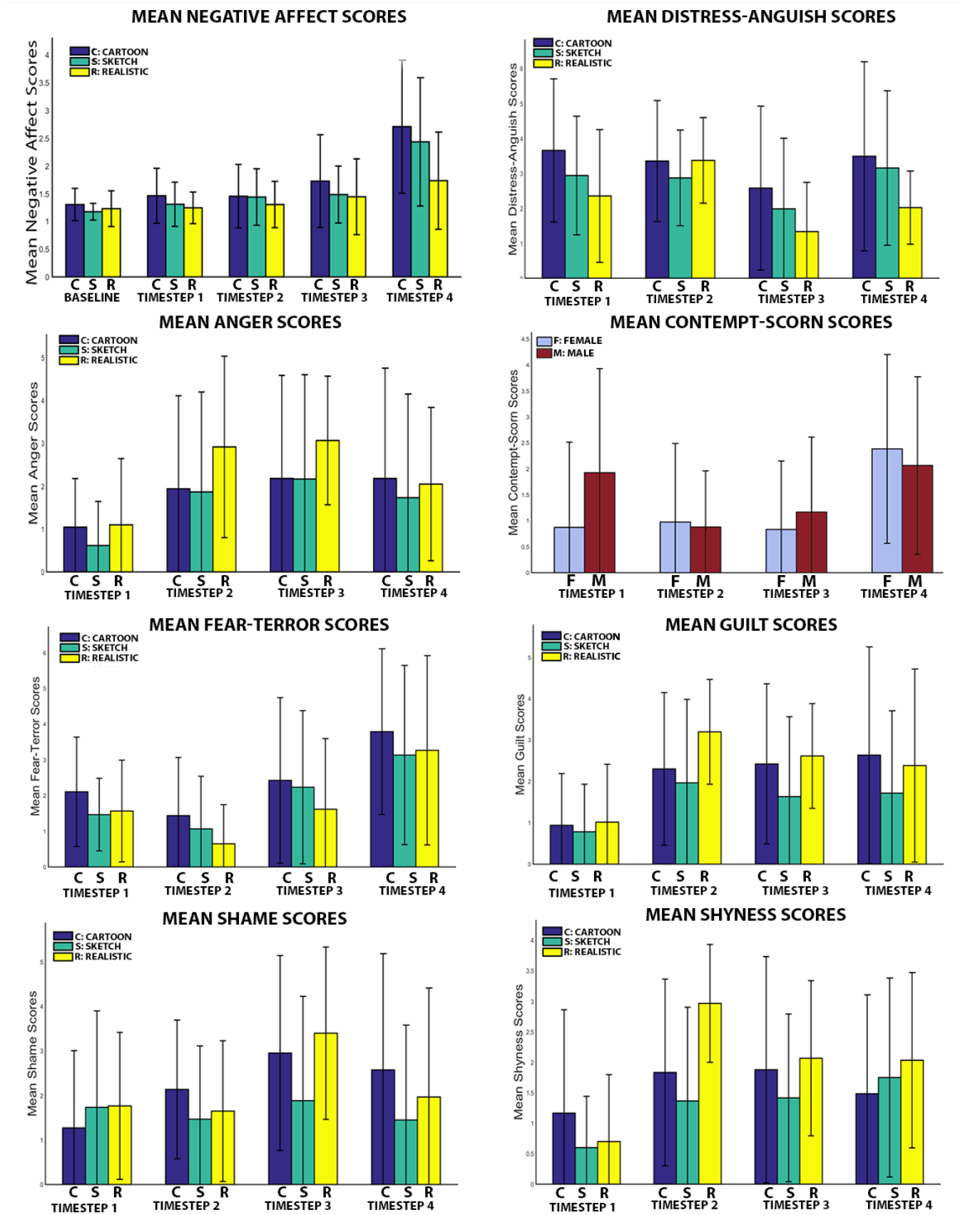

Fig. 6. Graphs showing Mean Scores on Negative Affect (a), and DES Dimensions of Distress-Anguish (b), Anger (c), Contempt-Scorn (d), Fear-Terror (e), Guilt (f), Shame (g), Shyness (h)

step2 (M=1.06, SD=1.44) $p < 0.001$, and after time-step3 (M=2.10, SD=2.15) $p < 0.001$ in all visual appearance conditions. Post-hoc comparisons also revealed that mean fear-terror scores were lowest after time-step2 and was significantly different than time-step1 $p = 0.006$ and time-step3 $p < 0.001$ (See Figure 6e).

In the dimension ***guilt***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of guilt scores across the different time-steps overall, $F(2.26, 126.48) = 16.68$, $p < 0.001$, $\eta^2 = 0.23$ (obs. power = 1.0), and reflected a pattern in the data very similar to that of mean anger scores. Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean guilt scores after time-step1 (M=0.91, SD=1.25) were the lowest overall, and significantly different as compared to after time-step2 (M=2.48, SD=1.79) $p < 0.001$, after time-step3 (M=2.23, SD=1.77) $p < 0.001$, and after time-step4 (M=2.26, SD=2.33) $p = 0.001$ corresponding to the gradual medical deterioration of the virtual patient Bob in all visual appearance conditions (See Figure 6f).

In the dimension ***shame***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of perceived shame scores across the different time-steps overall, $F(2.73, 153.01) = 10.67$, $p < 0.001$, $\eta^2 = 0.16$ (obs. power = 1.0). Pairwise post-hoc comparisons using Bonferroni adjusted alpha revealed that mean shame scores after time-step3 (M=2.75, SD=2.22) were the highest corresponding Bob's worsened medical condition, and was significantly different as compared to after time-step1 (M=1.58, SD=1.84) $p < 0.001$, after time-step2 (M=1.76, SD=1.59) $p < 0.001$, and after time-step4 (M=2.02, SD=2.41) $p = 0.01$ in all visual appearance conditions (See Figure 6g).

Finally, in the dimension ***shyness***, the mixed model ANOVA analysis revealed a significant main effect of sampling time of perceived shyness scores when interacting with Bob across the different time-steps overall $F(2.27, 127.37) = 15.26$, $p < 0.001$,$\eta^2 = 0.22$ (obs. power = 1.0), and a significant sampling time by condition interaction $F(4.55, 127.37) = 3.53$, $p = 0.007$, $\eta^2 = 0.11$ (obs. power = 0.88). In order to further examine the sampling time by visual appearance condition interaction on the mean perceived shyness scores, we conducted multiple block analyses (See Figure 6h). In condition C, pairwise post-hoc comparisons of mean shyness scores conducted via Bonferroni analysis did not reveal a significant difference. In condition S, pairwise comparisons revealed that the mean enjoyment-joy score was lowest after time-step1 (M=0.60, SD=0.84) and differed significantly from scores after time-step3 (M=1.42, SD=1.37) $p = 0.04$, and time-step4 (M=1.75, SD=1.63) $p = 0.01$. In condition R, pairwise comparisons showed that the mean perceived shyness scores were similarly lowest after time-step1 (M=0.70, SD=1.09), and differed significantly from scores after time-step2 (M=2.96, SD=.96) $p < 0.001$, time-step3 (M=2.06, SD=1.27) $p = 0.001$, and time-step4 (M=2.03, SD=1.44) $p = 0.001$. Time-step2 was the highest and also significantly differed from time-step4 $p = 0.21$. Mean perceived shyness scores were not significant different between conditions C, S, and R after time-step1 or time-step3 or time-step4. However, Tukey-HSD post-hoc analysis revealed that mean perceived shyness scores at time-step2 were significantly higher in the visually realistic condition (M=2.96, SD=0.96) than the cartoon condition (M=1.83, SD=1.53) $p = 0.027$, and the scores in the sketch condition that appeared in the middle (M=1.36, SD=1.54) $p = 0.001$.

5.1.4 Social Presence

We treated the set of scores on each of the responses to 5 questions on the social presence questionnaire [1] with a 2 x 3 two-way ANOVA analysis. The two independent variables were the gender and virtual human rendering condition that the participant experienced. Out of a total of 5 questions, we found a significant effect of condition on two of them. In response to the question "To what extent did you feel that Bob was sentient, conscious, and alive?" the ANOVA analysis revealed a significant main effect of condition, $F(2, 34) = 3.535$, $p = 0.041$, $\eta^2 = 0.13$. Post-hoc Tukey HSD revealed that participants in the Realistic condition (M=5.81, SD=0.43) perceived Bob as a sentient, conscious and alive significantly higher than participants in the Cartoon condition (M=4.25, SD=0.41), $p = 0.012$. In response to the question To what extent did you feel Bob was only a computerized image and not a real person?, the ANOVA analysis revealed a significant main effect of condition, $F(2, 34) = 3.47$, $p = 0.048$, $\eta^2 = 0.11$. Post-hoc Tukey HSD revealed that participants in condition S (M=4.0, SD=0.45) responded to this question with a significant higher (more like computer generated) than participants in condition R (M=2.54, SD=0.48) $p = 0.039$. The scores for the Cartoon condition (M=3.41, SD=0.46) on this question fell between the Realistic and Sketch conditions.

5.1.5 Relevant Performance Variables

Based on the log files of the interactions of the participants with Bob in the RRTS, we examined a number of performance related variables namely the average time in minutes spent with the Bob, average questions asked, and the average instrument interactions in measuring Bob's vital signs. These numerical dependent variables were treated with a 2 x 3 two way ANOVA analyses with gender (2 levels) and condition (3 levels) as independent variables. With respect to the average time spent with Bob, the ANOVA analysis revealed a significant main effect of gender $F(1, 56) = 5.10$, $p = 0.028$, $\eta^2 = 0.053$. Overall, female participants (M=9.75, SD=1.94) spent significantly more time on average with Bob than male participants (M=8.56, SD=1.31). The ANOVA analysis did not reveal any significant differences in mean instrument interactions or mean number of questions asked.

5.2 Qualitative Results

In order to assess the perceived differences between Realistic, Cartoon and Sketch conditions, at the end of the study we asked participants to report their overall impressions of interacting with the virtual patient Bob. In response to the question, ***"what did you like most about Bob?"*** Participants in the Realistic condition mentioned that Bob, "seemed and felt like a real patient", "seemed realistic and interactive, immersive, and felt very real", and "Bob interacts and answers promptly." Whereas, participants in the Cartoon condition said, "he could have more facial expressions", "he was trying to look at the bright side", "he looked cartoonish." Participants in the Sketch condition mentioned, "he was patient", "his voice made him sound real." In response to the question, ***"what did you like least about Bob?"*** Participants in the Realistic condition said, "friendly old guy under other circumstances, but he seemed to appear more and more agitated, distressed and unhealthy", "was weary and weak, his personality seemed very grumpy and kind at the same time", "seemed understanding at first, but then he started acting rude as you can see that he was experiencing more pain." Participants in the Cartoon condition said that Bob, "was not so cheerful." Whereas, participants in the Sketch condition said, "I can't think of Bob as someone in the flesh," and "can't think of him as someone exists." In response to the question, ***"how would you describe Bob's personality and appearance?"*** Participants in the Realistic condition said, "he followed me with his eyes", "his negative attitude", "he was rude and curt when answering questions", and "his appearance was a bit disconcerting." Participants in the Cartoon condition were more ambivalent, "he is pleasant, not annoying at all", "he is trying to be cheerful", "seemed like a kind person." Whereas, participants in the Sketch condition said he was, "old, weak and slim," and "old and sick."

6 Discussion

Our empirical evaluation investigated the effects of visually realistic human like appearance versus two types of stylized (Cartoon like and Charcoal Sketch like) rendering of virtual humans in evoking emotional responses in simulated inter-personal experiences. We also examined the role of gender in this interaction, as we wanted to investigate if males and females responded differently to the visual appearance conditions of the virtual human. From the PR to NPR rendering continuum [20], we implemented the Visually Realistic (R), Sketch (S) and Cartoon (C) like appearance conditions for a virtual patient "Bob in our RRTS that served as a rich experiment platform together with a failure to rescue scenario for evaluating emotional reactions of partici-

pants to the different levels of fidelity of the virtual human appearance in the context of a stressful emotional inter-personal experience.

Objective skin EDA results show that for all the three conditions, male participants exhibited significantly higher levels of arousal than female participants. Over the years, researchers have discovered considerable variability in their findings with respect to SCR and gender [5]. The findings in the research literature may be varying due to lack of standard measurement devices and metrics, varying size of the participant pool and different task loads. We hope to explore the gender differences in greater depth in future experiments towards validating our findings. Overall, male EDA levels were high through all the time-steps of the interaction with Bob. However, female EDA arousal levels seem to have decreased significantly from training session through to the last time-step of interaction with Bob, corresponding to the medical deterioration scenario. In previous work, we have not found any evidence for difference overall in male and female EDA data due to physiological factors [34] or in response to an emotion contagion stimuli in the virtual world [26, 31, 13]. Perhaps male participants in our simulation found the deterioration of Bob more intriguing and stimulating, as compared to female participants, potentially due to differences in game play experiences in males and females. We did not provide a game play experience survey in this experiment, and we aim to gather data in this metric in future studies. Female EDA levels were highest in the beginning but seemed to drop over time, which could be attributed to the novelty factor of the virtual reality simulation and decreases in subsequent time-steps corresponding to the medical deterioration of Bob.

The differences in visual appearance fidelity seems to have had an impact on subjective levels of negative affect. Overall, negative affect levels in participants rose from baseline through time-step4 in all three visual appearance conditions. The overall levels of negative affect are lower in the visually realistic condition, but seem to increase gradually from one time-step to the next. Whereas, in the cartoon and sketch conditions, negative affect levels seems to be slightly elevated and dramatically increases in time-step4. Since animation fidelity was not altered between the visual appearance conditions of the virtual patient, participants in the higher visual fidelity realistic condition may have perceived idiosyncrasies and nuances in Bobs expression more readily than in the cartoon and sketch like appearance condition and could be related to differing expectations of users in the visually realistic and non-realistic conditions.

A granular examination of the subjective responses to the ten DES dimensions of positive and negative emotion revealed that participants responded similarly with respect to some of the emotional dimensions and differently with respect to others, depending on the visual appearance conditions experienced. We found that regardless of which visual appearance condition the participants experienced, positive emotions such as interest-excitement and surprise seem to have gradually decreased and conversely negative emotions such as anger, fear-terror and guilt seem to have increased from the first time-step to the last corresponding to the medical deterioration and discomfort of Bob as time wore on. Similar to the skin EDA data, the higher levels of positive affect in the beginning of the simulation may be due to the novelty of the virtual reality simulation and may have subsequently decreased corresponding to the worsening state of Bob in further time-steps. In some emotional constructs such as enjoyment-joy, in time-steps pertaining to the visual distress of Bob, participants in the visually realistic condition exhibited significantly lower levels of enjoyment than participants in the cartoon and sketch conditions. Likewise, in other social emotional constructs such as shyness, in critical time-steps during which Bob deteriorated participants exhibited greater levels of shyness in the visually realistic condition than in the non-realistic cartoon and sketch conditions. This shows that despite critical negative emotional reactions being decreased in the visually realistic condition, participants still perceived "Bob" to a large extent as a human like conversational partner and exhibited a high degree of shame and shyness in the critical time-steps of the interaction as compared to the NPR conditions.

Based on the social presence results and self-reports, participants in the visually realistic condition perceived Bob as a real person, immersive and interactive, and like a real patient. They also perceived Bob with a complex and diverse personality, than in the sketch condition. Whereas, participants in the cartoon condition seem to have perceived Bob as lacking facial expression, some perceived him as rude, while other as friendly. Similar to the results of Zibrek and McDonnell [40], one novel find is that visual realism seems to have an impact on participants' perceptions of the virtual humans personality in inter-personal simulations. Realistic Bob was initially perceived as friendly, however in the later time-steps majority of participants consistently described him as rude, grumpy, curt and unfriendly corresponding to his deterioration in medical and emotional state. Taking the quantitative and qualitative results together, it appears that visual realism may allow users to perceive more subtle channels of information that can alter users' expectations and may enable them to perceive virtual humans with diverse personality traits and characteristics.

7 Conclusions

To our knowledge, this research is one of the first in empirically examining the role of visually realistic versus stylized appearances of a virtual human in a dynamic virtual reality simulation on the emotional responses of users. Our primary research question investigates if visual appearance fidelity of virtual humans in inter-personal simulations affect the emotional reactions of participants. Contrary to our expectations, we found that in our simulated affective inter-personal scenario surprisingly participants in the visually realistic condition exhibited the least negative affect corresponding to the deterioration of virtual patient "Bob" as compared to the Cartoon and Sketch shaded conditions, especially in the last time-step of the interaction. When examined at a granular level, this same pattern of results were found in the negative emotional dimensions of DES. Negative emotional responses in the realistic condition seem to vary a lot across the time-steps corresponding to the patient deterioration, and is lower in the last time-step as compared to the cartoon condition. With all other factors being the same such as facial expressions, gender, speech and behavior of the virtual patient in the different rendering conditions, the emotional reactions of the participants attests to the Uncanny Valley effect in that perhaps users are more inclined to notice the subtle behavioral cues such as gestural nuances and idiosyncrasies more readily in the visually realistic condition than the other NPR conditions, which in turn may have an effect in suppressing the critical expected emotional reactions of the participants in a simulated affective training scenario.

The unexpected diminished emotional responses of the participants in the critical times of a simulated inter-personal situation with virtual humans for training can also have a detrimental effect on the pedagogical benefits of the simulation as emotion and cognition are connected [29, 30]. Our findings have important implications to the design of virtual humans in complex medical training and social simulations. On one hand our results suggest that if animation and behavioral realism are lacking then perhaps a less visually realistic appearing virtual human may evoke a strong critical emotional reaction pertaining to the content and scenario of the simulated inter-personal simulation. On the other hand, our study conveys that visual realism is still important for eliciting appropriate social emotional constructs such as shyness and shame. More generally, our study suggests that virtual human appearance can affect not only critical emotional reactions in affective inter-personal training scenarios, but also users perceptions of personality and social characteristic of the virtual interlocutors. The results of this study have implications to the Uncanny Valley effect [21, 20] that we notice with virtual entities in movies and virtual simulations. It has been suggested that abstractions and stylized rendering can alter our expectation of virtual entities, so that we are not detracted by the lack of human idiosyncrasies and subtle expressive cues. Our research shows that interactions with stylized virtual humans in virtual reality systems can also potentially suppress levels of emotional bonding and alter affective reactions to the virtual humans in simulated inter-personal trainers. Diminished responses in critical affective dimensions may potentially negatively impact higher level functions such as learning, rapport, empathy, and trust.

8 Future Work

One of the limitations of our study is that in the non-realistic conditions, the virtual environment was not rendered according to the same cartoon or sketch algorithm. Instead, we strived to maintain consistency in the look and interactions of participants with the environment, medical instruments, and EHR across all three virtual human rendering conditions (R, C, and S). Even though none of the participants in the non-realistic conditions reported any perceived discrepancy in appearance between the virtual human and the environment, future studies will compare the results of the emotional responses in the non-visually realistic appearance conditions of the virtual human alone, to conditions in which the virtual human as well as the environment are rendered in a similar manner. Future work will also include an empirical evaluation of the effects of visual realism on participants' perception of sociality of the agents, such as persuasion, rapport formation and trust in simulated task-oriented encounters in inter-personal situations.

References

[1] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29(7):819–833, 2003.

[2] J. N. Bailenson, K. R. Swinth, C. L. Hoyt, S. Persky, A. Dimov, and J. Blascovich. 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*, 14(4):379–393, 2005.

[3] J. C. Campbell, M. J. Hays, M. Core, M. Birch, M. Bosack, and R. E. Clark. Interpersonal and Leadership Skills : Using Virtual Humans to Teach New Officers Interpersonal and Leadership Skills : Using Virtual Humans to Teach New Officers. (11358):1–11, 2011.

[4] H. Chaturvedi, N. D. Newsome, and S. V. Babu. An evaluation of virtual human appearance fidelity on user's positive and negative affect in human-virtual human interaction. In *Virtual Reality (VR), 2015 IEEE*, pages 163–164. IEEE, 2015.

[5] S. B. Daily, D. Meyers, S. Darnell, T. Roy, and M. T. James. Understanding privacy and trust issues in a classroom affective computing system deployment. In *Distributed, Ambient, and Pervasive Interactions*, pages 414–423. Springer, 2013.

[6] M. E. Dawson, A. M. Schell, and D. L. Filion. 7 the electrodermal system. *Handbook of psychophysiology*, pages 200–223, 2000.

[7] M. Dehghani, P. Carnevale, and J. Gratch. Interpersonal effects of expressed anger and sorrow in morally charged negotiation. *Judgment and Decision*, 9(2):104–113, 2014.

[8] D. DeVault, R. Artstein, G. Benn, and T. Dey. SimSensei kiosk: a virtual human interviewer for healthcare decision support. *Proceedings of the*, (1):1061–1068, 2014.

[9] S. D'Mello and J. Gratch. Creative expression of emotions in virtual humans. In *Proceedings of the 4th International Conference*, pages 336–338, 2009.

[10] L. C. Dukes, J. Bertrand, M. Gupta, R. Armstrong, T. Fasolino, S. V. Babu, and L. F. Hodges. A multitasking healthcare simulation for training nurses to recognize patient deterioration. In *Proceedings of Carolina Women in Computing 2012 (CWIC 2012)*, 2012.

[11] Q. Dunsworth and R. K. Atkinson. Fostering multimedia learning of science: Exploring the role of an animated agents image. *Computers & Education*, 49(3):677–690, Nov. 2007.

[12] P. M. Ellis and J. J. Bryson. The significance of textures for affective interfaces. In *Intelligent Virtual Agents*, pages 394–404. Springer, 2005.

[13] M. Grimshaw, C. A. Lindley, and L. E. Nacke. Sound and immersion in the first-person shooter: mixed measurement of the player's sonic experience. 2008.

[14] J. Healey and R. Picard. Digital processing of affective signals. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3749–3752. IEEE, 1998.

[15] J. Hodgins, S. Jorg, C. O'Sullivan, S. I. Park, and M. Mahler. The saliency of anomalies in animated human characters. *ACM Transactions on Applied Perception*, 7(4):1–14, Jul 2010.

[16] J. Hyde, E. J. Carter, S. Kiesler, and J. K. Hodgins. Assessing naturalness and emotional intensity: a perceptual study of animated facial motion. In *Proceedings of the ACM Symposium on Applied Perception*, pages 15–22. ACM, 2014.

[17] S. Levy. Why tom hanks is less than human. *Newsweek*, (September):18, 2004.

[18] K. F. MacDorman, R. D. Green, C.-C. Ho, and C. T. Koch. Too real for comfort? uncanny responses to computer generated faces. *Computers in human behavior*, 25(3):695–710, 2009.

[19] R. L. Mandryk, D. Mould, and H. Li. Evaluation of emotional response to non-photorealistic images. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering - NPAR '11*, page 7. ACM Press, 2011.

[20] R. McDonnell, M. Breidt, and H. H. Bülthoff. Render me real? *ACM Transactions on Graphics*, 31(4):1–11, July 2012.

[21] M. Mori. Bukimi no tani [the uncanny valley]. pages 33–35, 1970.

[22] M. Obaid, I. Damian, F. Kistler, B. Endrass, J. Wagner, and E. André. Cultural behaviors of virtual agents in an augmented reality environment. In *Intelligent Virtual Agents*, pages 412–418. Springer, 2012.

[23] X. Pan, D. Banakou, and M. Slater. Computer based video and virtual environments in the study of the role of emotions in moral behavior. *Affective Computing and Intelligent*, pages 52–61, 2011.

[24] M.-Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on*, 57(5):1243–1252, 2010.

[25] H. Prendinger, J. Mori, and M. Ishizuka. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International Journal of Human-Computer Studies*, 62(2):231–245, Feb. 2005.

[26] H. Prendinger, J. Mori, and M. Ishizuka. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2):231–245, 2005.

[27] C. Qu, W. Brinkman, and Y. Ling. Conversations with a virtual human: Synthetic emotions and human responses. *Computers in Human*, pages 58–68, 2014.

[28] L. Ring, D. Utami, and T. Bickmore. The right agent for the job? In *Intelligent Virtual Agents*, pages 374–384. Springer, 2014.

[29] A. Robb, R. Kopper, R. Ambani, F. Qayyum, D. Lind, L.-M. Su, and B. Lok. Leveraging virtual humans to effectively prepare learners for stressful interpersonal experiences. *IEEE transactions on visualization and computer graphics*, 19(4):662–70, Apr. 2013.

[30] D. L. Roter, R. M. Frankel, J. a. Hall, and D. Sluyter. The expression of emotion through nonverbal behavior in medical visits. Mechanisms and outcomes. *Journal of general internal medicine*, 21 Suppl 1:S28–34, Jan. 2006.

[31] M. Slater, P. Khanna, J. Mortensen, and I. Yu. Visual realism enhances realistic response in an immersive virtual environment. *Computer Graphics and Applications, IEEE*, 29(3):76–84, 2009.

[32] J. Tsai, E. Bowring, S. Marsella, W. Wood, and M. Tambe. A study of emotional contagion with virtual characters. In *Intelligent Virtual Agents*, pages 81–88. Springer, 2012.

[33] J. van der Schalk, A. Fischer, B. Doosje, D. Wigboldus, S. Hawk, M. Rotteveel, and U. Hess. Convergent and divergent responses to emotional displays of ingroup and outgroup. *Emotion*, 11(2):286, 2011.

[34] P. Venables and D. Mitchell. The effects of age, sex and time of testing on skin conductance activity. *Biological psychology*, 43(2):87–101, 1996.

[35] V. Vinayagamoorthy, M. Gillies, and A. Steed. Building expression into virtual characters. 2006.

[36] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[37] A. C. Wellerdiek, M. Breidt, M. N. Geuss, S. Streuber, U. Kloos, M. J. Black, and B. J. Mohler. Perception of strength and power of realistic male characters. 2015.

[38] Y. Wu, S. V. Babu, R. Armstrong, J. W. Bertrand, J. Luo, T. Roy, S. B. Daily, L. C. Dukes, L. F. Hodges, and T. Fasolino. Effects of virtual human animation on emotion contagion in simulated inter-personal experiences. In *IEEE transactions on visualization and computer graphics*, volume 20, pages 626–635, 2014.

[39] C. Zanbaka and A. Ulinski. Social responses to virtual humans. *Conference on HCI*, pages 1561–1570, 2007.

[40] K. Zibrek and R. McDonnell. Does render style affect perception of personality in virtual humans? In *Proceedings of the ACM Symposium on Applied Perception*, pages 111–115. ACM, 2014.