Leveraging Virtual Humans to Effectively Prepare Learners for Stressful Interpersonal Experiences

Andrew Robb, Regis Kopper, Ravi Ambani, Farda Qayyum, David Lind, Li-Ming Su, and Benjamin Lok, *Member, IEEE*

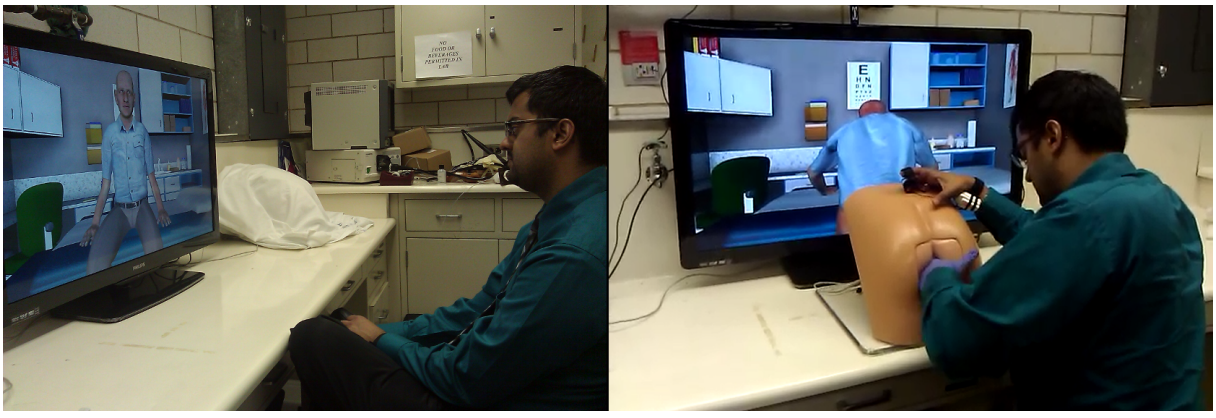


Fig. 1. On the left: a user interviews a virtual human about his urinary difficulties. On the right: a user examines a virtual human's prostate for abnormalities. A skin conductance sensor is visible on the user's right hand.

Abstract—Stressful interpersonal experiences can be difficult to prepare for. Virtual humans may be leveraged to allow learners to safely gain exposure to stressful interpersonal experiences. In this paper we present a between-subjects study exploring how the presence of a virtual human affected learners while practicing a stressful interpersonal experience. Twenty-six fourth-year medical students practiced performing a prostate exam on a prostate exam simulator. Participants in the experimental condition examined a simulator augmented with a virtual human. Other participants examined a standard unaugmented simulator. Participants' reactions were assessed using self-reported, behavioral, and physiological metrics. Participants who examined the virtual human experienced significantly more stress, measured via skin conductance. Participants' stress was correlated with previous experience performing real prostate exams; participants who had performed more real prostate exams were more likely to experience stress while examining the virtual human. Participants who examined the virtual human showed signs of greater engagement; non-stressed participants performed better prostate exams while stressed participants treated the virtual human more realistically. Results indicated that stress evoked by virtual humans is linked to similar previous real-world stressful experiences, implying that learners' real-world experience must be taken into account when using virtual humans to prepare them for stressful interpersonal experiences.

Index Terms—Virtual/digital characters, mixed reality, training, user studies

1 Introduction

Virtual humans have been shown to evoke stress in humans. Slater and Pertaub have shown humans experience stress when inflicting pain on a virtual human and when practicing public speaking in front of a critical virtual audience [22] [16]. Virtual humans have also been successfully applied to various types of interpersonal skill training including negotiation[24], medical interviews [6], and team training [18]. Virtual humans' ability to cause stress and their usefulness for interpersonal skills training suggests that virtual humans can be leveraged to safely prepare learners for stressful interpersonal experiences. In this paper, we explore how the presence of a virtual human affects learners' behavior during a stressful interpersonal experience.

We conducted a between-subjects study (n = 26) designed to explore how a virtual human affects learners while simulating a stressful interpersonal experience. The stressful interpersonal experience selected for the study was performing a simulated prostate exam. Performing a prostate exam is a commonly simulated part of medical education. Performing a prostate exam is a stressful interpersonal experience containing both procedural and social elements. Procedural elements include using proper examination technique and examining the entire prostate; social elements include explaining the exam to the patient, reassuring the patient during the exam, and coping with discomfort induced by the intimate nature of the exam. Standard prostate exam simulators focus on teaching procedural skills and provide little to no preparation for social components of the exam. Simulating interpersonal components of the exam using virtual humans may cause students to experience more realistic stress when practicing on prostate exam simulators, as the intimate interpersonal nature of the exam is a significant source of stress [17]. Experiencing interpersonal stress while practicing prostate exams could lead to reduced stress while performing real prostate exams; phobia [21, 15] and post-traumatic stress disorder (PTSD) [20] research has demonstrated that exposure to stressful virtual experiences translates to reduced stress in similar real-world experiences.

Prostate cancer is the fifth leading cause of death in American men;

- *Andrew Robb is with the University of Florida. E-mail: arobb@ufl.edu.*
- *Regis Kopper is with Duke University. E-mail: regis.kopper@duke.edu.*
- *Ravi Ambani is with Drexel University. E-mail: rna26@drexel.edu.*
- *Farda Qayyum is with Drexel University. E-mail: fq23@drexel.edu.*
- *David Lind is with Drexel University. E-mail: david.lind@drexelmed.edu.*
- *Li-Ming Su is with the University of Florida. E-mail: li-ming.su@urology.ufl.edu.*
- *Benjamin Lok is with the University of Florida. E-mail: lok@cise.ufl.edu.*

Manuscript received 13 September 2012; accepted 10 January 2013; posted online 16 March 2013; mailed on 16 May 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

the American Cancer Society estimates that there were 241,740 new cases and 28,170 deaths from prostate cancer in 2011. The American Cancer Society recommends that men ages 50 years and older should consider undergoing annual screening for prostate cancer, including a prostate exam and a prostate specific antigen (PSA) blood test [23]. As prostate exams are an important component of prostate cancer screening, improving prostate exam performance would likely lead to improved patient outcomes. Reducing learners' stress has the potential to significantly improve the effectiveness of prostate exams, as stress has been linked to poorer task performance [13].

Despite the importance of early screening for prostate cancer, medical students receive few to zero opportunities to practice performing a real prostate exam in a safe environment. Many hospitals use human volunteers to train medical students how to perform a prostate exam, but students rarely receive more than one session. Medical students can practice procedural skills using prostate exam simulators, but, as was previously discussed, standard prostate exam simulators provide students with little to no preparation for the stressful interpersonal components of performing a prostate exam.

Thirteen participants performed a prostate exam using a simulator augmented with a virtual patient; these participants also interviewed the virtual patient about his urinary difficulties before performing the exam and explained their diagnosis to him after the exam. The remaining thirteen participants performed a prostate exam using a standard simulator. Participants received feedback about their exam performance in an after-action review. Surveys, behavioral measures, and skin conductance were measured as dependent variables.

2 Related Work

2.1 Affective Responses Towards Virtual Humans

Virtual humans have been shown to elicit a wide range of affective responses in humans, including stress [22], fear of public speaking [16], racial-bias[19], paranoia [4], and self-disclosure [7].

Slater et al. recreated Stanley Milgram's famous shock experiment using a virtual human as the shock recipient. Slater found that participants experienced stress when commanded to deliver extreme shocks to the virtual human, though the degree of stress was smaller than that observed by Milgram when using a real human subject [22].

Pertaub et al. conducted a study in which he compared the effects of three different virtual audiences on participants' anxiety while delivering a speech. He found that the negative virtual audience evoked anxiety among participants [16].

Rossen et al. compared participants' interactions with a virtual patient in which only the skin tone of the patient was varied. He found that participants' real-world biases were expressed towards a dark-skinned patient [19].

In summary, prior work has shown that virtual humans elicit affective responses similar to those elicited by real human beings. Slater and Pertaub's demonstrations that virtual humans can evoke stress supports the hypothesis that participants will experience more realistic stress levels when examining a physical exam simulator augmented with a virtual human. Rossen's findings concerning racial bias indicates that participants' reactions to virtual humans are mediated by prior experience with real humans.

2.2 Virtual Humans Use In Interpersonal Skills Training

Virtual humans have been applied to many aspects of interpersonal skills training, including medical interviews [6], breast exams [8], and negotiations [24].

Johnsen et al. studied the validity of using virtual humans to prepare learners for interpersonal experiences. He found that performance during a simulated medical interview with a virtual human predicted performance in a simulated medical interview with a human actor [6].

Virtual humans have also been applied to training procedural skills. Kotranza et al. created a mixed reality human (a virtual human that incorporates physical components) used to practice performing a breast exam. Kotranza found that participants treated the MRH as a social actor, using interpersonal touch to comfort and reassure the patient. Kotranza compared exams performed on the MRH to exams performed on a human actor. He found that participants performed equivalent exams, demonstrating the usability of MRH's for physical exam training. Kotranza did not evaluate the MRH compared against a standard breast exam simulator [8].

Traum et al. developed and piloted a virtual human system for non-team negotiation, supporting multiple virtual humans each operating under their own agenda. Traum found that participants were able to carry out negotiations which succeeded or failed based on participants' use of proper negotiation strategies [24].

In summary, virtual humans have been successfully applied in a wide range of interpersonal skills training applications. The use of virtual humans as a training tool for medical education has been well established. Kotranza demonstrated that participants performed equivalent breast exams on a MRH and a human actor, validating virtual humans' usefulness for simulating intimate exams. However, he did not explore if participants responded differently to a MRH and an unaugmented breast exam simulator. Traum showed that virtual humans can be used for training complex interpersonal experiences.

2.3 Prostate Exam Simulation

Most research concerning simulation of intimate exams has focused on the breast exam. Comparatively little research has been directed towards simulating the prostate exam.

Research related to simulating prostate exams has focused on detecting exam performance metrics and improving tactile fidelity. Balkissoon et al. described a simulator which incorporates eight force sensors. These sensors were used to observe the areas participants examined and how much pressure they applied. Balkissoon found that experts performed significantly better exams than novices [2].

Low-Beer et al. describes a simulator which allows for visual observation of the exam. Prostate exams normally cannot be observed because they take place inside of the body. Low-Beer modified a simulator and added cameras which captured the exam while the exam was being performed [11] .

Kowalik et al. describes a high-fidelity simulator capable of simulation of multiple different conditions. Kowalik used inflatable balloons to simulate the elasticity of a human prostate and allow for dynamic simulation of nodules. Kowalik's simulator included four force sensors capable of detecting the pressure applied by participants [9].

We collaborated with Balkissoon in the development of our simulator. We extended upon Balkisson's work by adding additional sensors and developing a system to deliver feedback to users about their exam.

3 System

3.1 Virtual Human Used For Prostate Exam Simulation

The Prostate Virtual Interactive Character (Patrick) was developed in conjunction with an experienced urologist. Patrick simulated a 65-year-old man complaining of difficulty when urinating. Patrick's difficulty urinating was caused by an enlarged prostate, a common problem among older men which is not dangerous.

Participants' interactions with Patrick were split into three stages (see Figure 2). Participants interviewed Patrick about his symptoms, medical history, medications, and his family history. On average, participants interviewed Patrick for four and a half minutes. Once they had completed their interview, participants explained to Patrick that they needed to perform a prostate exam. Participants then performed a prostate exam, which lasted twenty to thirty seconds, and afterward explained their diagnosis and preferred treatment plan to Patrick. The entire interaction lasted approximately seven minutes.

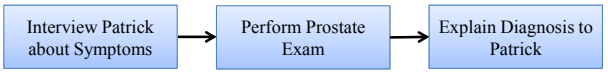


Fig. 2. Participants' interactions with Patrick were split into three stages: gathering information from Patrick about his symptoms, performing a prostate exam, and explaining the diagnosis to Patrick.

3.1.1 Visual Appearance

Participants interacted with Patrick via a 40" wide-screen LCD-TV placed at eye-level in front of participants (shown in Figure 1). Participants could see Patrick seated on an examination table in a doctor's office. Participants interacted with Patrick via spoken voice. Patrick responded with audio pre-recorded by a 63-year-old voice talent.

3.1.2 Conversational Model

Patrick could respond to questions using one of 136 responses related to his symptoms, medical and family history, medications, and personal information. Patrick could also respond to unanticipated questions using one of 30 generic responses. Generic responses included phrases like "Yes", "Yeah, I guess so", "No", and "Not that I can think of". Patrick was built with the assistance of an experienced urologist who provided input on the questions participants would ask and appropriate responses. Patrick was able to provide satisfactory answers to approximately 90% of questions asked by students.

3.1.3 Wizard-of-Oz

A Wizard-of-Oz system was used to allow participants to speak naturally with Patrick and eliminate confounding effects associated with speech recognition and understanding. An operator listened to participants and selected the most appropriate response for any given question or statement. Participants were not aware that an operator was controlling the virtual human.

On average, Patrick responded to participants within one second. The operator could often determine an appropriate response before participants finished speaking. Operator performance was improved by grouping responses categorically. Responses were selected via a touchpad; using a touchpad eliminated the noise associated with a mouse click.

3.1.4 Stressful Interpersonal Moments

Patrick simulated stressful interpersonal moments present when performing a prostate exam on a real patient. These stressful moments included explaining the need to perform a prostate exam, seeing the patient disrobed, and performing the exam.

Participants had to explain to Patrick that they needed to perform a prostate exam. Participants explained this after they had finished their interview. Upon learning that a prostate exam was needed, Patrick responded by saying "Shoot. I figured you'd have to do one. Can't say I've been looking forward to it. [pause] So how do we do this?" Participants then explained what would happen during the prostate exam.

Multiple participants commented that Patrick's reluctance to being examined felt realistic. One participant described Patrick, saying he seemed "like any other normal patient I might encounter in an office setting. Calm, but still reluctant to be examined (like any other normal man about to get a rectal exam)." Another participant described him as "slightly worried and hesitant".

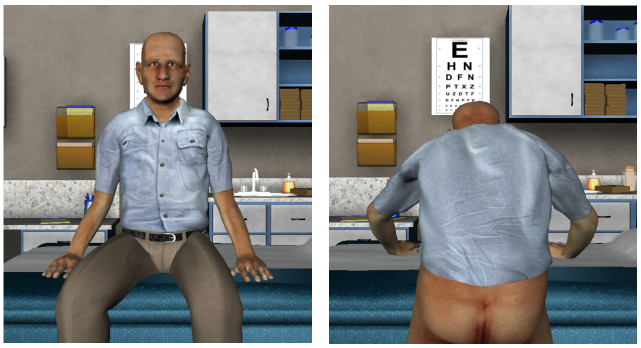
Patrick asked participants to turn around while he undressed. Once Patrick was undressed, he informed participants that he was ready. Participants could see that Patrick was undressed from the waist down, as a patient would be during a real exam. (See Figure 3)

3.2 Prostate Exam Simulation

Prostate exams are performed by inserting a gloved and lubricated index finger into a patient's rectum and palpating the entire surface of the prostate, which lies immediately inside and below the rectum. Examiners palpate the prostate by sweeping their finger back and forth across the prostate's surface. While palpating, examiners must assess the size of the prostate and feel for hard nodules that may indicate the presence of prostate cancer. Nodules indicative of prostate cancer can be missed if the examiner does not palpate the entire surface of the prostate or does not use sufficient pressure during palpation.

3.2.1 Stressful Components of the Prostate Exam

Several factors contribute to novices' anxiety about performing prostate exams. Pugh surveyed 1st year medical students about sources of anxiety when performing prostate exam. Pugh found that common sources of anxiety included: the intimate nature of the exam, causing harm or pain to the patient, the embarrassment of the patient, and sanitation issues. Of the students Pugh surveyed, only 1.7% reported that they felt no anxiety about performing prostate exams. [17]



(a) Interview (b) Exam

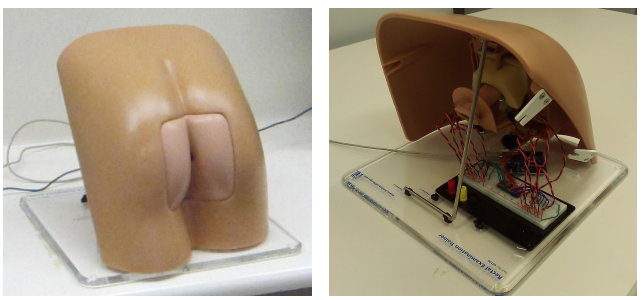
Fig. 3. Patrick before and after undressing. Patrick's exposed buttock could be seen after he had undressed and turn around.

3.2.2 Construction of the Simulator

Prostate exam simulators consist of a plastic shell simulating the buttocks, a flexible tube simulating the rectum, and a rubber module representing the prostate. Different interchangeable prostate modules are used to simulate different clinical conditions, including normal prostates, enlarged prostates, and prostates with nodular cancer. Commercially available trainers do not include any sensors capable of assessing how well an exam was performed.

We extended a prostate exam simulator purchased from Limbs and Things. Twelve force sensitive resistors were added to the prostate module which represented an enlarged prostate. Each force sensor had a sensing area of 0.3". Sensors could detect forces ranging between 0.1N and 10N. Pressure applied during an exam rarely exceeded 8.0 N. Sensors were sampled at 30 Hz.

A subject matter expert evaluated the simulator and determined that the simulator was capable of detecting exam performance. Prostate exams are performed by sweeping the index finger back and forth across the surface of the prostate. This sweeping motion can be leveraged to reduce the number of sensors required. So long as sensors are within one finger's width (0.5") of each other, the intervening space can be assumed to have been examined when two sensors have been activated.



(a) Front (b) Rear

Fig. 4. The prostate exam simulator seen from the front and rear. Sensors are attached to a prostate module underneath

3.2.3 Creation of an Expert Exam Used for Assessment

The prostate exam simulator must be calibrated before use. Calibration is performed in two stages: creation of a model expert exam and calculation of static pressure.

A model expert exam is used to assess student performance. An expert creates a model expert exam by examining the simulator several

times. The maximum pressure applied to each sensor during the exam is recorded. The maximum pressures recorded during each exam are then averaged together to create a model of the pressure which should be applied to a sensor. This is done for each sensor.

Sensors occasionally detect some static pressure even when an exam is not being performed. This static pressure is caused by the interior of the simulator. The static pressure of each sensor must be calculated at the beginning of each exam, as the static pressure can change after the simulator has been examined. Static pressures typically ranged from 0.0 N to 4.0 N. The static pressure never exceeded the amount used in the model expert exam.

3.2.4 Feedback Provided About Exam Performance

Participants' exams were recorded and used to provide feedback about their exam performance in an after-action review. Participants were shown a visualization which replayed their exam (See Figure 5). The visualization divided the prostate into twelve regions, each containing a single sensor. The color of each region changed based on the amount of pressure being applied. Regions were outlined in green once participants had applied the amount of pressure in the model expert exam.

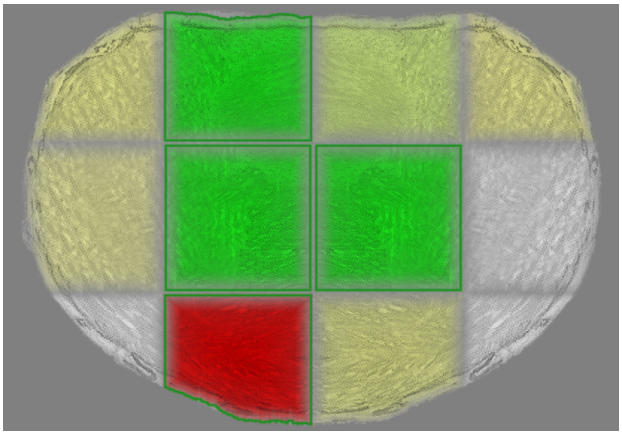


Fig. 5. A snapshot of one participant's feedback. Pale yellow regions are receiving little to no pressure, green regions are receiving sufficient pressure, and red regions are receiving more pressure than required. Regions which are outlined have been examined sufficiently.

4 Study Design

We conducted a study to determine how the presence of a virtual patient impacted participants' behavior while examining a prostate exam simulator.

Two types of prostate exam simulator were used during the study (Figure 6): a standard simulator and an augmented simulator. The standard simulator contained twelve pressure sensors and was visually indistinguishable from other common prostate exam simulators. The augmented simulator extended the standard simulator using Patrick, a virtual human representing a patient complaining of difficulty urinating.

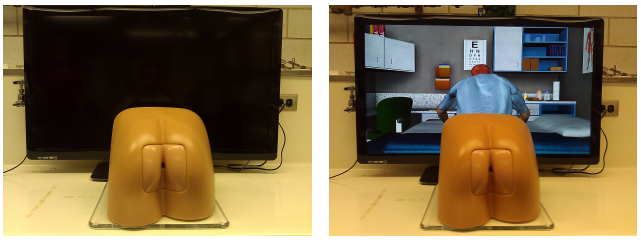
4.1 Hypotheses

We hypothesize that participants will experience more stress while examining a prostate exam simulator augmented with a virtual human than while examining a standard unaugmented prostate exam simulator.

4.2 Population

Twenty-seven fourth-year medical students at Drexel University participated in this study. One student's results were removed due to technical difficulties with the virtual patient. The final number of participants was twenty-six, with eleven men (n = 26).

65% of students had performed at least one prostate exam on a real human prior to the study. 89% of students had performed at least one prostate exam on either a real human or a simulator prior to the study. All students had received lectures concerning how to perform prostate exams.



(a) Standard Simulator (b) Augmented Simulator

Fig. 6. The standard and augmented prostate exam simulators. The same physical prostate exam module was used in both the standard simulator and the augmented simulator.

4.3 Procedure

Participants performed three prostate exams during the study: a baseline exam, an experimental exam, and an assessment exam (see Figure 7). The baseline exam was used to assess participant's existing skill at performing prostate exams. The experimental exam was used to explore the effect of a virtual human on participant behavior while performing a prostate exam. The assessment exam was used assess whether participants had improved their performance after examining the simulator and receiving feedback.

Participants used the standard simulator when performing the baseline exam and the assessment exam. The simulator used during the experimental exam was varied between conditions.

Participants received feedback about their experimental and assessment exams. Feedback took place after the exam had been completed.

4.4 Conditions

Participants were divided into two conditions. Each condition followed the same procedure during the baseline exam and assessment exam; each condition followed a different procedure during the experimental exam.

Participants in the first condition examined the augmented simulator during the experimental exam. We refer to this condition as the *augmented condition*. Participants in the augmented condition interviewed Patrick about his symptoms, performed an exam, and explained their diagnosis to Patrick after the exam.

Participants in the second condition examined the standard simulator during the experimental exam. We refer to this condition as the *standard condition*. Participants in the standard condition only performed an exam; they did not see or speak with a virtual human.

4.5 Evaluation Methods

Three forms of evaluation assessed the effect of a virtual human on participants' behavior while performing a prostate exam: self-report metrics (surveys), behavioral metrics (prostate exam performance), and physiological metrics (skin conductance).

4.5.1 Self-Report Metrics

Participants were surveyed about their comfort and confidence with performing prostate exams using the following questions:

1. "How comfortable are you with the prospect of performing a prostate exam on a patient?"
2. "How confident are you in your ability to perform a prostate exam on a patient?"

Participants rated their comfort/confidence using an asymmetric six point scale, ranging from "extremely uncomfortable", "very uncomfortable", "uncomfortable", "somewhat uncomfortable", "comfortable" to "very comfortable". This scale was used by Pugh when

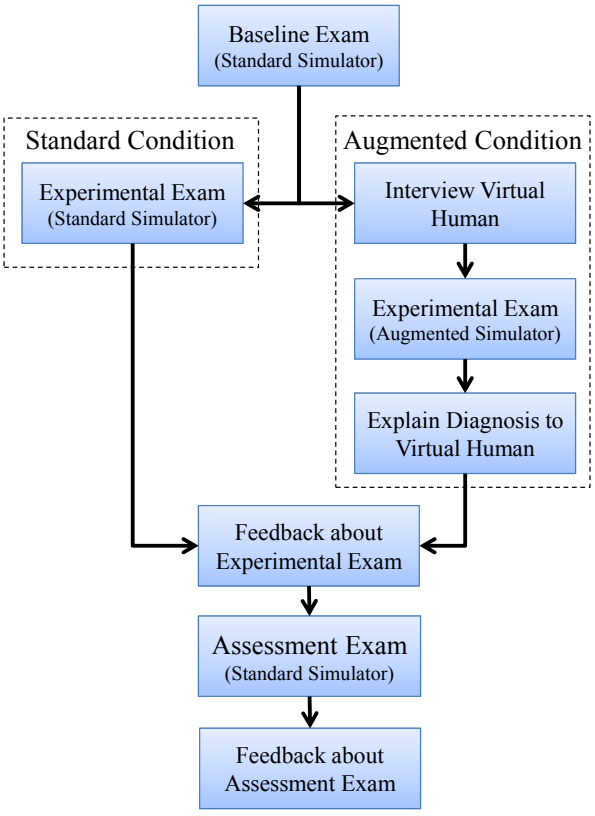


Fig. 7. Participants were split into two conditions. All participants performed the same baseline and assessment exam; participants performed different experimental exams based on condition. Participants in the augmented condition spoke with a virtual human before, during, and after the augmented exam.

assessing students' anxiety related to performing prostate exams. Pugh used discomfort as an indicator of anxiety [17].

Participants who interacted with the augmented simulator completed an additional survey at the end of the study. This study contained questions about Patrick's realism, a co-presence question, and Bailenson's social presence questionnaire [1]. Co-presence is defined as "having a feeling that one is in the same place as the other participants, and that one is collaborating with real people." [3] Social presence is defined as the "sense of being with another." [10].

Participants also completed a follow-up survey one week after the study; the follow-up survey assessed participants' social anxiety and fear of negative evaluation use the SAD and FNE scales [25].

4.5.2 Behavioral Metrics

Behavioral metrics provide an objective means of comparing participants. Participants' exam performance was used as a behavioral metric. Exam performance was broken into two components: percentage of the prostate examined (coverage) and time spent performing the exam (duration). Coverage was measured relative to the model expert exam; a region was considered covered if a participant examined the region using at least 75% of the pressure applied by the expert.

4.5.3 Physiological Metrics

Physiological metrics provide objective data which can be used to compare participants' reactions. Skin conductance was used as a measure of participants' physiological arousal during the study. Studies have shown that mental stress can increase participant's physiological arousal [5]. Participants' SCR was collected at 30 Hz using wireless Shimmer SCR sensors placed on the index and middle finger of participants' non-dominant hand (visible in Figure 1). Participants' SCR was collected between the initial and final survey.

5 Results

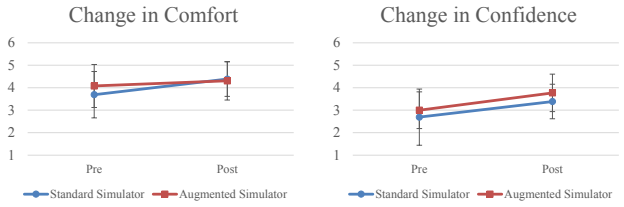
Thirteen participants were randomly assigned to each condition. No significant differences between the standard and augmented conditions were observed in gender distribution, age, number of prior prostate exams performed, or self-reported comfort or confidence.

Mixed design ANOVAs were used to explore the relationship between the presence of a virtual human and participants' pre- and post-interaction behavior. Condition was used as the between-subject factor and self-reported metrics and exam performance were used as within-subject factors.

5.1 Self-Reported Metrics

Participant comfort only improved when a virtual human was not present. A mixed design ANOVA explored if comfort changed pre- and post-interaction, and if changes varied between conditions (see Figure 8). Comfort increased by 0.46 (out of six) ($F_{1,24} = 18.78, p < .0001$) between pre- and post-interactions. However, an interaction effect was observed based on condition. A pairwise comparison showed that the change in comfort was significant in the standard condition ($\mu_{standard} = 0.69, p < .0001$) but not in the augmented condition ($p = .139$). No selection bias was observed between conditions ($F_{1,24} = .205, p = .655$).

Participant confidence improved irrespective of the presence of a virtual human. A mixed design ANOVA explored if confidence changed pre- and post-interaction, and if changes varied between conditions (see Figure 8). Confidence increased by 0.73 (out of six) ($F_{1,24} = 36.71, p < 0.0001$) between pre- and post-interactions. No interaction effects were observed based on condition. No selection bias was observed between conditions ($F_{1,24} = .102, p < 0.753$).



(a) Change in Comfort (b) Change in Confidence

Fig. 8. The presence of a virtual patient affected changes in participant comfort. Participants' comfort increased when examining the standard simulator, but did not change when examining the augmented simulator. The presence of a virtual patient did not affect changes in participant confidence. Comfort and Confidence were rated on six point scales. See the section 4.5.1 for more details.

5.2 Behavioral Metrics

Participants' exam performance improved only when a virtual patient was present. A mixed design ANOVA assessed if participants' exam performance changed from the baseline to the assessment exams, and if changes varied between conditions. Participants examined 9.1% more of the prostate in the assessment exam than the baseline exam ($F_{1,23} = 6.647, p < 0.017$). No significant interaction effect was observed between conditions, but a Bonferroni pairwise comparison revealed that the difference between baseline and assessment exams was only significant for participants in the augmented condition ($p_{standard} = 0.252, p_{augmented} < 0.05$). No selection bias was observed in baseline exams between conditions ($p = 0.374$). Participants' exam performance is shown in Figure 9.

Participants' performance during the experimental exam was not included in the analysis due to an observed confound: participants in the standard condition performed cursory experimental exams because they had already examined the prostate and were acquainted with its

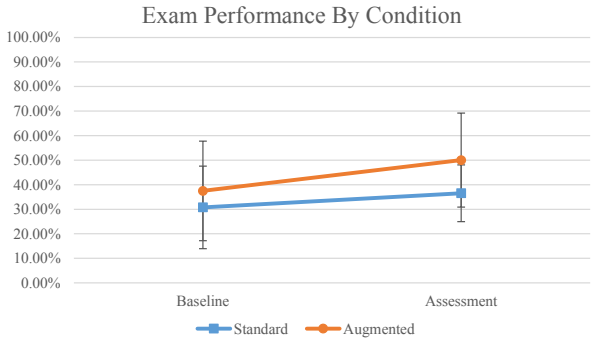


Fig. 9. Percentage of the prostate examined at the beginning and end of the study. Participants' exam performance improved only in the Augmented condition.

findings. Participants did not perform a cursory assessment exam; this was likely related to having received feedback about their performance. Participants in the augmented condition did not perform a cursory experimental exam, which may have been related to the presence of a virtual human.

5.3 Physiological Metrics

Participants in the augmented condition exhibited higher arousal. No significant differences were seen in participants' initial SCR. Significant differences ($p < 0.05$) were seen between conditions in participants average and peak SCR, with participants in the augmented condition exhibiting higher average and peak values. Significant differences ($p < 0.05$) were also seen in the variance of participants' SCR, indicating that participants' SCR varied more strongly in the augmented simulator condition. See Table 1 for values.

Table 1. Participants' Average, Peak, and Variance SCR by condition. Significant differences were seen between all measures shown.

Condition	Average (μS)	Peak (μS)	Variance (μS)
Standard	247.2($\pm$9.9)	259.0($\pm$11.2)	13.4($\pm$8.8)
Augmented	273.0($\pm$37.4)	299.8($\pm$51.6)	161.8($\pm$188.3)

Four participants' SCR were not collected due to technical malfunctions. One participant's data was discarded after he reported running to the study; heightened physical activity is known to alter SCR. Twenty-one valid SCR readings were collected, eleven in the standard simulator condition and ten in the augmented simulator condition.

Analysis indicated that one participant in the standard simulator condition was an outlier. We define an outlier as the following: values outside of the interquartile range by a factor of 1.5 are mild outliers; values outside of the interquartile range by a factor of 3.0 are extreme outliers [14]. This participant's average SCR was a mild outlier, and his peak SCR and variance in SCR were extreme outliers. His SCR data was excluded from our analysis.

5.3.1 Skin Conductance Response Sub-Groups

Two sub-groups were identified in the augmented condition, based on SCR. A visual analysis of SCR in the augmented condition indicated the existence of two sub-groups based on SCR (see Figure 11). Linear regressions of participants' SCR indicated the presence of two significantly different sub-groups in the augmented condition. A significant difference ($p < 0.001$) was observed in the slope of the linear regression of these two sub-groups. The SCR of 50% participants in the augmented condition trended upwards (average slope = 0.0063); the SCR of the remaining 50% of participants remained level (average slope = -0.00026). We refer to participants exhibiting an upward trend in SCR as *strong-response* participants. We refer to participants exhibiting a level response as *weak-response* participants. Each sub-group contained five participants.

All participants in the standard condition exhibited a level response, with an average slope of -0.00020.

5.4 Analysis of Differences Between Sub-Groups

Physiological, behavioral, and self-report metrics were re-examined in light of the strong and weak response sub-groups to identify if there were any other differences between sub-groups. No significant differences between strong and weak response sub-groups were seen in self-reported comfort or confidence, social anxiety, fear of negative evaluation, gender, or age.

5.4.1 Self-Reported Differences Between Sub-Groups

No significant differences between sub-groups were observed in ratings of either pre- or post-interaction comfort/confidence

Strong-response participants had performed more real exams. A significant difference ($p < 0.05$) was observed between sub-groups in the number of exams participants had performed on human patients. On average, strong-response participants had performed three exams on a real human, while weak-response participants had performed 0.6 exams on a real human. All strong-response participants had performed at least one exam on a real human.

Strong-response participants reported higher co-presence. A significant difference between sub-groups ($\mu_{strong} = 4.2, \mu_{weak} = 3.4, p < .05$) was observed in participants' answers to the co-presence question. Every strong-response participant selected a positive response ("slightly-agree" or "agree"), while weak-response participants were split between positive and negative responses ("slightly disagree" and "slightly agree").

Strong-response participants generally reported a higher sense of social presence. Strong-response participants' reported a sense of social presence that was greater than or equal to weak-response participants for each question on Bailenson's social presence questionnaire [1], however, no statistically significant differences were observed.

5.4.2 Behavioral Differences Between Sub-Groups

Only weak-response participants showed significant improvements in exam performance A mixed design ANOVA assessed if exam performance was affected by participants' arousal level. The ANOVA found a main effect indicating that participants' exam performance improved from pre- to post-interaction ($F_{2,19} = 8.634, p < 0.01$). An interaction effect was observed based on sub-group ($F_{2,19} = 3.605, p < 0.05$). A Bonferroni pairwise comparison revealed that only weak-response participants' performance improved significantly ($p_{weak} < 0.05, p_{strong} = 0.208, p_{standard} = 0.240$). No selection bias was observed between weak-response and strong-response participants' baseline exams ($p = 0.648$). Figure 10 shows changes in exam performance for all participants.

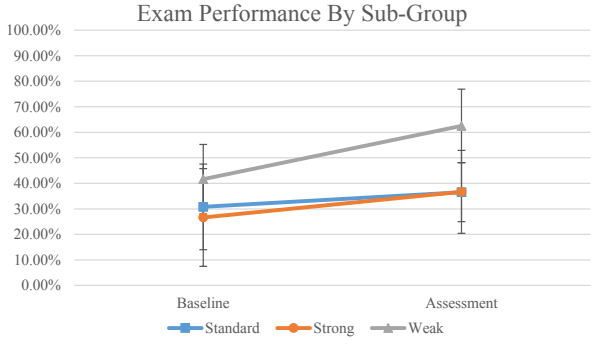


Fig. 10. Participants' prostate exam performance during the baseline and assessment exams. Significant differences between the baseline and assessment exams were only seen in weak-response participants. Despite the difference in average baseline performance, no significant differences in the baseline exam were seen between sub-groups or conditions.

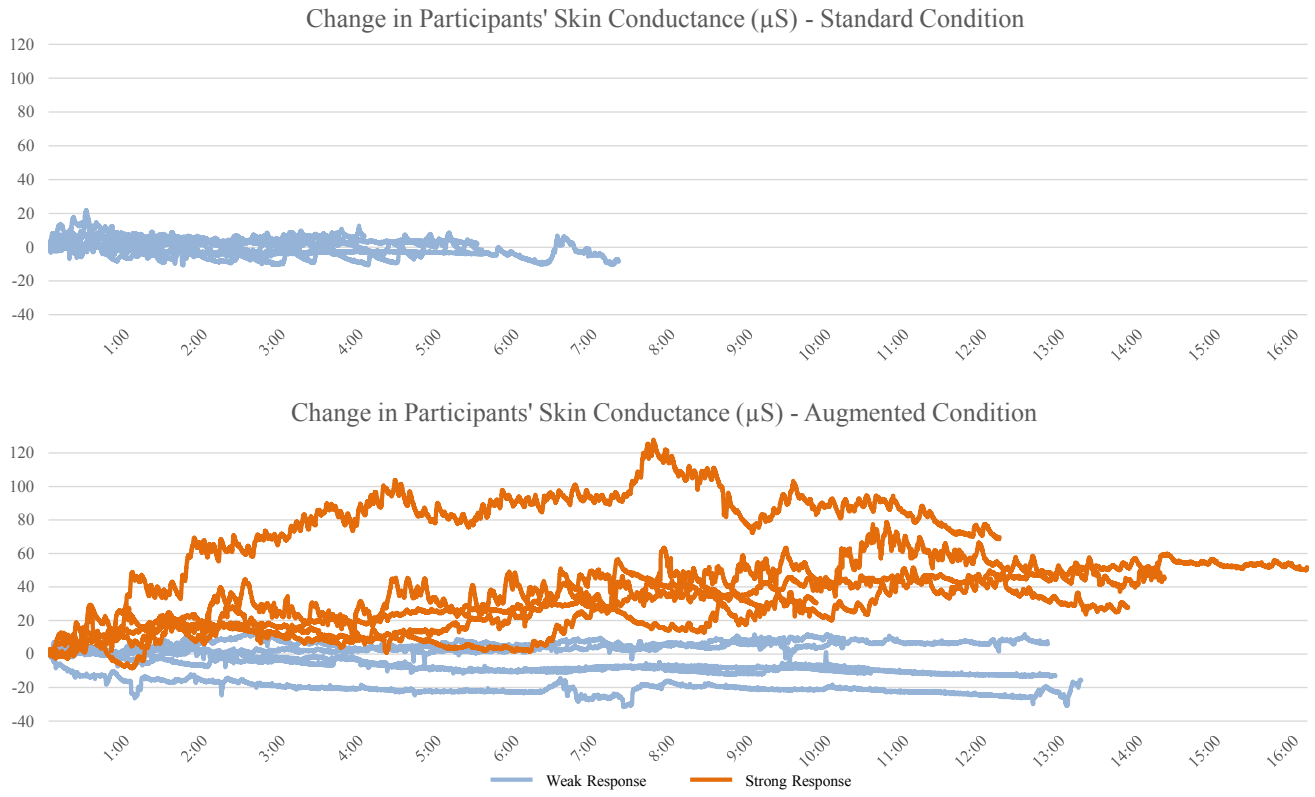


Fig. 11. Participants' SCR during the study. A weak SCR generally remained flat throughout the study. A strong SCR generally trended upwards throughout the study. Participants took more time in the augmented condition because of the interview with Patrick.

Strong-response participants talked with the virtual human more. A significant difference ($p < 0.05$) between sub-groups was observed in the number of questions addressed to the virtual human. Strong-response participants spoke to the virtual human 48.6($\pm$8.96) times, while weak-response participants spoke to the virtual human 36.6($\pm$7.54) times. A significant difference ($p < 0.05$) was also seen in the length of interviews. On average, strong-response participants spoke with the virtual human for 5:55 ($\pm$ 1:40) minutes, while weak-response participants spoke for 3:50 ($\pm$ 1:06) minutes.

5.4.3 Physiological Differences Between Sub-Groups

Strong differences in average, peak, and variance in SCR were also seen between sub-groups. Strong significant differences were seen between sub-groups in average SCR ($p < 0.01$), peak SCR ($p < 0.001$), and variance in SCR ($p < 0.01$). No significant differences were seen in the initial SCR, though a trend was observed ($p < 0.1$) indicating that the strong response sub-group's initial SCR was higher. See Table 2 for values.

Table 2. Participants' Average, Peak, and Variance SCR by condition and sub-group. Significant differences were seen between strong-response participants and other participants.

Condition	Average (μS)	Peak (μS)	Variance (μS)
Standard	247.2($\pm$9.9)	259.0($\pm$11.2)	13.4($\pm$8.8)
Weak	244.4($\pm$8.9)	258.5($\pm$9.4)	15.5($\pm$6.6)
Strong	301.6($\pm$31.9)	341.1($\pm$40.3)	308.2($\pm$161.8)

6 Discussion

The presence of a virtual human improved participants' performance in one of two ways: either participants' procedural skills improved more after receiving feedback, or participants exhibited more social engagement while performing the prostate exam. The response participants demonstrated appears to be related to their physiological arousal and their previous real-world experience performing prostate exams.

6.1 Interpreting Participants' Physiological Arousal

Two explanations for participants' arousal seem likely: participants' arousal may be indicative of either stress or feelings of presence. Both stress and feelings of presence are known to cause arousal measurable by skin conductance [5, 12]. Research has shown that performing a prostate exam is a stressful experience for medical students. The correlation between real-world experience and arousal suggests that participants' may have been primed by stressful real world experiences, causing them to experience stress when examining the virtual patient. This response is consistent with phobia research [21, 15] and PTSD research [20], which have established that stressful memories can cause stress in virtual environments.

Alternatively, participants' arousal may have been caused by stronger feelings of presence. This argument is strengthened by noting that strong-response participants showed signs of stronger social engagement: they asked the virtual human more questions, spent more time talking with the virtual human, and reported stronger feelings of co-presence. It may be that participants' arousal is simply another indication of stronger feelings of presence. This would be consistent with Meehan's use of skin conductance as a physiological measure of feelings of presence [12].

Both explanations seem plausible. However, there is no reason to assume that stress and presence are mutually exclusive. They may in fact be complementary. Meehan's research linking skin conductance and presence took place in a stressful virtual environment; he found that people experienced stronger feelings of presence in stressful virtual environments. Similarly, Slater's work relating skin conductance and presence involved stressful social situations. Slater recreated Stanley Milgram's shock experiment using virtual humans; Milgram's original experiment caused participants to experience crippling

stress. Participants in Slater's experiment experienced similar, though greatly reduced, stress. Slater found that participants experienced a greater sense of presence in the more stressful condition [22]. In light of Meehan's and Slater's work, and the stressful nature of performing a prostate exam, it seems likely that strong-response participants' arousal indicates stress accompanied by stronger feelings of presence.

6.2 Leveraging Participants' Varied Responses to Virtual Humans

While weak-response participants were not physiologically aroused by the virtual human, they were still affected by the virtual human. Weak-response participants' exam performance improved significantly between their baseline and assessment exams; no improvements were observed for strong-response or standard-condition participants. The presence of a virtual human may have caused weak-response participants' to feel that the stakes were higher when performing the prostate exam, leading them to try harder and pay more attention to the feedback they received. This in turn led to successfully internalizing their feedback, which improved their exam performance.

Strong-response participants' exam performance did not improve, but they did show signs of greater social engagement. Strong-response participants reported a higher sense of co-presence, spent more time talking with the virtual human, and asked him more questions. Strong-response participants' failure to improve their exam performance suggests that they failed to internalize their feedback. It seems likely that the stress caused by the virtual human interfered with strong-response participants' ability to successfully internalize their feedback.

It is intriguing that the presence of a virtual human improved different elements of the prostate exam for participants with low and high arousal. Each sub-group outperformed the other in one of the two main components of the prostate exam: examining the prostate and communicating with the patient. Weak-response participants performed a better exam, yet communicated worse with the patient. Strong-response participants performed worse prostate exams, but talked with the patient more. The presence of a virtual human enabled weak-response participants to internalize feedback about their exam performance, but results indicate that they experienced weaker feelings of presence. The presence of a virtual human caused strong-response participants to behave more realistically when interacting with the patient, but failed to help them improve their exam. This divergence in responses suggests that care should be taken when using virtual humans for training, as they may produce divergent effects in participants who perceive the virtual human differently.

6.3 Importance of Simulating Interpersonal Components of Physical Exams

Preparing learners for stressful interpersonal experiences using purely procedural simulators may have undesired effects. Participants reported how confident they were in their ability to perform a prostate exam and their comfort with the prospect of performing a prostate exam. These two measures can be mapped to the procedural and interpersonal elements of the prostate exam. It is not surprising that both conditions reported similar increases in confidence; both conditions practiced the same procedural skills and received similar feedback. It is surprising that standard participants reported an increase in comfort, while augmented participants did not. This may indicate that standard procedural simulators falsely inflate learners' sense of preparation for interpersonal components of physical exams. Learners may not properly assess the difficulty of interpersonal components of physical exams when interpersonal components are not simulated. Incorporating interpersonal elements via a virtual human may prevent learners from developing a false sense of comfort with the interpersonal components of physical exams.

6.4 Limitations

Our ability to interpret this data is limited by the small number of participants in both sub-groups. The presence of two sub-groups was unanticipated, or more participants would have been recruited. Our small n was exacerbated by technical difficulties which resulted in losing SCR data from several participants. However, the existence of consistent significant differences between groups at these sample sizes is encouraging. It is likely that larger sample sizes will only strengthen our conclusions and allow more detailed exploration of how the presence of a virtual human impacts learner performance.

The SCR of the outlier in the standard-condition showed similarities with the SCR of strong-response participants. It is possible that this participant represents a small contingent of students who experience stress while interacting with standard prostate exam simulators. This participant had previously performed two real prostate exams, indicating he may have experienced stress while performing a real prostate exam. It is also possible that his high arousal was due to something unrelated to the prostate exam. A larger study would help clarify if a small contingent of students exist who experience stress when examining standard prostate exam simulators.

7 Conclusions and Future Work

In this paper we described a study which explored how the presence of a virtual human affected learners while practicing a stressful interpersonal experience. Participants only experienced stress when examining a prostate exam simulator augmented with a virtual human and participants who had performed more exams were more likely to be stressed. Unstressed participants improved their prostate exam performance while stressed participants treated the virtual human more realistically.

Stressed participants failed to internalize and apply feedback about their exam performance. It is important to consider this failure when preparing learners for stressful interpersonal experiences. When possible, learners should be taught procedural skills before being exposed to a stressful real-world experience. This may help learners internalize feedback and prepare them to perform better during stressful real-world experiences.

The presence of a virtual human produced divergent effects in participants. It is important to consider that virtual humans may produce unexpected or undesirable effects in some participants. In this case, both prominent responses were positive. With careful consideration, it is likely that most effects caused by a virtual human can be leveraged, even if they are not necessarily desired.

We are currently working to integrate this technology with medical education. Integrating virtual humans with medical education will open up opportunities to study the effects of long-term exposure to stressful interpersonal experiences involving virtual humans, as well as allowing medical educators to study the long-term effects of improved prostate exam training on patient safety.

Acknowledgments

The authors wish to thank Dr. Carla Pugh for her assistance in constructing our prostate simulator, and our participants for their time and effort. This work was supported in part by NSF Grant IIS-0803652 and the Cornelius F.J. Beukenkamp Endowment for Prostate Cancer Research.

References

[1] J. Bailenson and J. Blascovich. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 2003.

[2] R. Balkissoon, K. Blossfield, L. Salud, D. Ford, and C. Pugh. Lost in translation: unfolding medical students' misconceptions of how to perform a clinical digital rectal examination. *American Journal of Surgery*, 197(4):525–32, Apr. 2009.

[3] J. Casanueva and E. Blake. The effects of avatars on co-presence in a collaborative virtual environment. *Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, 2001.

[4] D. Freeman, M. Slater, P. E. Bebbington, P. a. Garety, E. Kuipers, D. Fowler, A. Met, C. M. Read, J. Jordan, and V. Vinayagamoorthy. Can virtual reality be used to investigate persecutory ideation? *The Journal of Nervous and Mental Disease*, 191(8):509–14, Aug. 2003.

[5] S. C. Jacobs, R. Friedman, J. D. Parker, G. H. Tofler, a. H. Jimenez, J. E. Muller, H. Benson, and P. H. Stone. Use of skin conductance changes

during mental stress testing as an index of autonomic arousal in cardiovascular research. *American Heart Journal*, 128(6 Pt 1):1170–7, Dec. 1994.

[6] K. Johnsen, A. Raij, A. Stevens, D. Lind, and B. Lok. The validity of a virtual human experience for interpersonal skills education. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1049–1058. ACM, 2007.

[7] S.-H. Kang and J. Gratch. The effect of avatar realism of virtual humans on self-disclosure in anonymous social interactions. *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, page 3781, 2010.

[8] A. Kotranza, B. Lok, A. Deladisma, C. Pugh, and D. Lind. Mixed reality humans: Evaluating behavior, usability, and acceptability. *Visualization and Computer Graphics, IEEE Transactions on*, 15(3):369–382, 2009.

[9] C. G. Kowalik, G. J. Gerling, a. J. Lee, W. C. Carson, J. Harper, C. a. Moskaluk, and T. L. Krupski. Construct validity in a high-fidelity prostate exam simulator. *Prostate cancer and prostatic diseases*, 15(1):63–9, Mar. 2012.

[10] E. Lansing and C. Harms. Toward a More Robust Theory and Measure of Social Presence : Review and Suggested Criteria. *Presence: Teleoperators & Virtual Environments*, 12(5):456–481, 2002.

[11] N. Low-Beer, T. Kinnison, S. Baillie, F. Bello, R. Kneebone, and J. Higham. Hidden practice revealed: using task analysis and novel simulator design to evaluate the teaching of digital rectal examination. *The American Journal of Surgery*, 201(1):46–53, 2011.

[12] M. Meehan and B. Insko. Physiological measures of presence in stressful virtual environments. *ACM Transactions on Graphics*, page 645, 2002.

[13] D. E. Mina Westman. The inverted-U relationship between stress and performance: A field study. *Work & Stress*, 10(2), 1999.

[14] G. P. Moore, D. S. and McCabe. *Introduction to the Practice of Statistics, 3rd ed*. W. H. Freeman, New York, 1999.

[15] T. D. Parsons and A. a. Rizzo. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: a meta-analysis. *Journal of behavior therapy and experimental psychiatry*, 39(3):250–61, Sept. 2008.

[16] D.-p. Pertaub, M. Slater, G. Street, L. Wce, and C. Barker. An Experiment on Public Speaking Anxiety in Response to Three Different Types. *Presence: Teleoperators & Virtual Environments*, 11(1):68–78, 2002.

[17] C. M. Pugh, K. B. Iannitelli, D. Rooney, and L. Salud. Use of mannequin-based simulation to decrease student anxiety prior to interacting with male teaching associates. *Teaching and learning in medicine*, 24(2):122–7, Jan. 2012.

[18] J. Rickel. Virtual humans for team training in virtual reality. *Proceedings of the Ninth International Conference on Artificial Intelligence*, (July):578–585, 1999.

[19] B. Rossen, K. Johnsen, and A. Deladisma. Virtual humans elicit skin-tone bias consistent with real-world skin-tone biases. *Intelligent Virtual Agents*, pages 237–244, 2008.

[20] B. O. Rothbaum, L. Hodges, R. Alarcon, D. Ready, F. Shahar, K. Graap, J. Pair, P. Hebert, D. Gotz, B. Wills, and D. Baltzell. Virtual reality exposure therapy for PTSD Vietnam Veterans: a case study. *Journal of traumatic stress*, 12(2):263–71, Apr. 1999.

[21] B. O. Rothbaum, L. Hodges, P. L. Anderson, L. Price, and S. Smith. Twelve-month follow-up of virtual reality and standard exposure therapies for the fear of flying. *Journal of Consulting and Clinical Psychology*, 70(2):428–432, 2002.

[22] M. Slater, A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, N. Pistrang, and M. V. Sanchez-Vives. A virtual reprise of the Stanley Milgram obedience experiments. *PloS one*, 1(1):e39, Jan. 2006.

[23] A. C. Society. Cancer Facts & Figures 2012. Technical report, 2012.

[24] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. *Intelligent Virtual Agents*, 2008.

[25] D. Watson and R. Friend. Measurement of social-evaluative anxiety. *Journal of consulting and clinical psychology*, 33(4):448–57, Aug. 1969.