Comparing Interpersonal Interactions with a Virtual Human to Those with a Real Human

Andrew B. Raij, Kyle Johnsen, Robert F. Dickerson, *Student Member*, *IEEE*, Benjamin C. Lok, *Member*, *IEEE*, Marc S. Cohen, Margaret Duerson, Rebecca Rainer Pauly, Amy O. Stevens, Peggy Wagner, and D. Scott Lind

Abstract—This paper provides key insights into the construction and evaluation of interpersonal simulators—systems that enable interpersonal interaction with virtual humans. Using an interpersonal simulator, two studies were conducted that compare interactions with a virtual human to interactions with a similar real human. The specific interpersonal scenario employed was that of a medical interview. Medical students interacted with either a virtual human simulating appendicitis or a real human pretending to have the same symptoms. In Study I $(n = 24)$, medical students elicited the same information from the virtual and real human, indicating that the content of the virtual and real interactions were similar. However, participants appeared less engaged and insincere with the virtual human. These behavioral differences likely stemmed from the virtual human's limited expressive behavior. Study II $(n = 58)$ explored participant behavior using new measures. Nonverbal behavior appeared to communicate lower interest and a poorer attitude toward the virtual human. Some subjective measures of participant behavior yielded contradictory results, highlighting the need for objective, physically-based measures in future studies.

Index Terms—Intelligent agents, virtual reality, human-centered computing, user interfaces, interaction styles, evaluation/methodology, computer graphics, medicine.

1 Introduction

The Virtual Patient system ([1], [2], [3]) allows medical students to practice a difficult interpersonal situation—the medical interview—through interaction with a virtual human. Just as flight simulators help pilots improve flight skills, *interpersonal simulators* like the Virtual Patient system have the potential to help users improve *interpersonal communication skills*. This paper explores the potential of interpersonal simulators by comparing interactions with a virtual patient, *a virtual human that simulates a patient*, to interactions with a standardized patient, *a real human that simulates a patient*.

- *A.B. Raij, K. Johnsen, and B.C. Lok are with the Department of Computer Science and Engineering, University of Florida, PO Box 116125, Gainesville, FL 32611-6120. E-mail: {raij, kjohnsen, lok}@cise.ufl.edu.*
- *R.F. Dickerson is with the Department of Computer Science, University of Virginia, 151 Engineer's Way, PO Box 400740, Charlottesville, VA 22904-4740. E-mail: rfd7a@cs.virginia.edu.*
- *M.S. Cohen is with the College of Medicine, University of Florida, VA Medical Center, PO Box 100247, Gainesville, FL 32610-0247. E-mail: cohenms@surgery.ufl.edu.*
- *M. Duerson is with the College of Medicine, University of Florida, PO Box 100281, Gainesville, FL 32610-0281. E-mail: MDuerson@dean.med.ufl.edu.*
- *R.R. Pauly is with the College of Medicine, University of Florida, PO Box 100277, Gainesville, FL 32610-0277. E-mail: PAULYRR@medicine.ufl.edu.*
- *A.O. Stevens is with the NF/SG VA Health System and College of Medicine, University of Florida, VA Medical Center, PO Box 100286, Gainesville, FL 32610-0286. E-mail: amy.stevens@med.va.gov.*
- *P. Wagner is with the School of Medicine, Medical College of Georgia, HB 3040, Augusta, GA 30912. E-mail: pwagner@mcg.edu.*
- *D.S. Lind is with the Department of Surgery, Medical College of Georgia, BB 4514, Augusta, GA 30912. E-mail: dlind@mcg.edu.*

Manuscript received 31 July 2006; revised 18 Oct. 2006; accepted 29 Nov. 2006; published online 22 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-0118-0706.
Digital Object Identifier no. 10.1109/TVCG.2007.1030.

Standardized patients are used extensively in medical schools worldwide. Next to seeing a real patient, they are the most effective way to train medical students on patient interaction. As the standardized patient interaction is a validated simulation of a real medical interview, it is the ideal gold standard to compare the virtual patient interaction to. To our knowledge, no other work has been published where an interpersonal simulator is formally compared to a validated real-world counterpart. This comparison is key to learning how to build and evaluate effective interpersonal simulators.

This paper describes two studies that compare standardized patient interactions to virtual patient interactions. Participants were medical students who interviewed either 1) a standardized patient trained to simulate the symptoms of appendicitis (Fig. 1a) or 2) a virtual human programmed to simulate the same symptoms (Fig. 1b). The interactions were then compared on the content of the interview, the behavior of participants, and the authenticity of the interaction.

Study I ($n = 24$, where n is the number of participants), presented at IEEE Virtual Reality 2006 [4], found that interactions with the standardized patient and virtual human were similar on gathering critical information from the patient and other content measures. Subtle differences were found on behaviors related to rapport with the patient. Participants appeared less engaged and insincere with the virtual human. Differences on rapport-building behaviors stemmed from the virtual human's limited expressiveness. Ultimately, Study I was limited because it did not sufficiently characterize participant behavior. Study I highlighted the need to develop new measures of behavior in interpersonal interactions.

Building on Study I, **Study II** $(n = 58)$ sought to 1) further characterize how behavior changes with virtual humans

Fig. 1. (a) A real interpersonal interaction and (b) an equivalent virtual interpersonal interaction. In the real interaction, a medical student interviews a real standardized patient. In the virtual interaction, the medical student interviews a virtual human.

using new measures and 2) strengthen the main findings of Study I.

Study II differed from Study I in that:

- New behavioral measures were added to characterize rapport with the virtual human and standardized patient.
- The system was integrated into a patient communication course. This gave access to a larger sample of the general medical student population. It also guaranteed that all students had the same experience level and training.
- The expressiveness of the virtual human was improved by recording the voice of a skilled standardized patient.

Study II strengthened Study I's results and also provided new insight into rapport-building behavior. Participants' nonverbal behavior communicated lower interest in the interaction and a poorer attitude toward the virtual human. Some new behavioral measures were too subjective to yield useful information. This highlighted the need for more objective, physically-based measures of human behavior in future studies.

Overall, the studies provide key insights into the **construction** and **evaluation** of effective interpersonal simulators.

Construction: If the content of the interaction is similar to its real-world counterpart, then an interpersonal scenario where information gathering is the main task can be effectively simulated. However, more complex scenarios that incorporate communication skills, like rapport-building, are more difficult to simulate because virtual humans are not as expressive as real humans. The expressiveness of virtual humans must be improved to elicit natural behavior from users.

Evaluation: Evaluating an interpersonal simulator objectively is difficult. More objective measures of interaction authenticity and participant behavior must be developed.

2 Previous Work

2.1 Effective Virtual Humans

Researchers have worked to establish the basis of effective virtual humans. Badler et al. [5] suggest virtual humans "should move or respond like a human" and "must exist, work, act, and react within a 3D virtual environment." Alessi and Huang [6] expand these rules for psychology applications. They suggest virtual humans should be social, emotionally expressive, and interactive. Virtual humans should "capture and assess a viewer and their emotional status, then translate this, taking into consideration cultural, educational, psychosocial, cognitive, emotional, and developmental aspects, and give an appropriate response that would potentially include speech, facial, and body emotional expression."

Thórisson and Cassell [7] agree that emotional expression is important, but nonverbal behaviors that support conversation, e.g., gesturing at objects and looking at the user to indicate attention, are more significant. In [8], Cassell et al. focus on modeling conversational interaction to create believable, functional virtual humans. Vinayagamoorthy et al. [9] concluded that 1) the behavioral and visual fidelity of virtual humans must be consistent, and 2) a virtual human's expressions should be appropriate for the application's context. In a later article [10], they review models for making virtual humans expressive. Nass et al. explored the affective power of computers and intelligent agents. Their work has shown that people ascribe human characteristics to computers, such as helpfulness, usability, and friendliness [11].

2.2 Human Behavior with Virtual Humans

There is growing evidence that people treat virtual and real people similarly. Pertaub et al. [12] noted that participants with a fear of public speaking reported similar anxieties when speaking to an audience of virtual people. Garau et al. [13] showed that people represented by avatars communicate better when the avatars employ realistic, task-appropriate eye-gaze.

Bailenson et al. have shown that people manage personal space similarly with real and virtual humans. People kept more distance from an embodied tutor than strangers [14], and more distance from embodied agents than inanimate virtual objects [15]. Female participants kept more distance from agents that maintained eye contact than with agents that did not.

Zanbaka et al. have shown that virtual entities (human or animal-like) can be as effective as real people at persuasion

[16]. In interactions with both real and virtual speakers, persuasion was stronger when participants listened to a speaker of the opposite-sex. In a different study, Zanbaka et al. [17] found that, as with real humans, the presence of a virtual human lowers performance on novel or complex tasks.

2.3 Virtual Human Applications

Virtual humans have been used in many applications. Thórisson's [18] interactive guide, Gandalf, gives solar system tours. USC's Institute for Creative Technologies created virtual experiences to train soldiers in interpersonal leadership [19]. Just VR [20] allows a medical trainee to work with a virtual assistant to assess and treat a virtual victim. Balcisoy et al. [21] created a system where users play chess against a virtual human augmented to the real world. The Human Modeling and Simulation Group at the University of Pennsylvania uses virtual humans for task analysis and assembly validation [22]. The Virtual Classroom [23] uses virtual teachers and students to assess attention and social anxiety disorders.

2.4 Virtual versus Real Experiences

Although little work directly compares real and virtual interpersonal scenarios, researchers have compared other virtual environments to their real counterparts. In the psychology domain, Emmelkamp et al. [24] compared the reactions of acrophobes (persons with a fear of heights) in virtual and real environments. The authors found that exposure therapy in the virtual environment was as effective as therapy in the real one. Rothbaum et al. [25] found similar results for treating the fear of flying. Experiencing a virtual airplane was just as effective as experiencing a real one in reducing flying anxiety.

Others have looked at human perception of real and virtual stimuli. Billger [26] examined the perception of color in virtual and real environments. Wuillemin et al. [27] looked at the perception of virtual and real spheres presented visually and with haptics. Virtual spheres presented visually were perceived as larger than real spheres of the same size.

Heldal et al. [28] studied collaboration in real and virtual environments. Participants collaborated on building a Rubik's cube in real or shared virtual environments. Performance in symmetric environments (e.g., both participants collaborating through an immersive projection system) approached performance in real environments. Performance in asymmetric environments (e.g., HMD versus immersive projection) was poorer.

Slater et al. [29] looked at the behavior of small groups in real and virtual environments. Participants viewed immersed peers as leaders in the virtual scenario, but not in the real one. Group accord was higher in the real environment.

Usoh et al. [30] examined participant responses on presence questionnaires after experiencing a real environment or a similar virtual environment. Participants indicated they felt just as present in the virtual environment as in the real environment. This surprising result shows that subjective questionnaires should not be used to compare different environments.

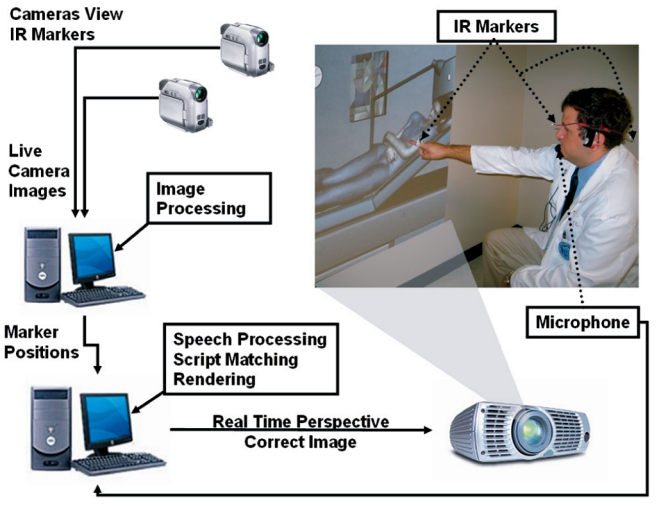

Fig. 2. System overview.

3 Virtual Patient System

The Virtual Patient system [1] (see Fig. 2) allows practice of medical interview skills. Students can gather the key facts of a patient's condition and arrive at a differential diagnosis, a list of conditions the patient may have. They also can practice communicating clearly with the patient and addressing their fears. These communication skills do not just improve with clinical experience. They should be taught and practiced [31].

As part of the studies, the system was installed in a real medical exam room. A virtual exam room was projected on a wall. DIANA, a virtual human with severe stomach pain, was in the virtual room. DIANA's appearance and responses are modeled after a real standardized patient, Maria, trained to exhibit severe stomach pain. Modeling DIANA after Maria allowed participants to interact with similar patients in the real and virtual experiences. Another virtual human, VIC, served as an instructor that tutors students on how to interact with DIANA. Commercial speech recognition software and a simple algorithm for parsing utterances [3] enabled talking to VIC and DIANA naturally within the scope of the scenario. The student's hand was tracked, allowing DIANA's pain to be localized with pointing gestures. The student's head was also tracked to render the scene from her perspective and allow the virtual human to maintain eye contact.

4 Study I: Design

One group of students (Group VH) interviewed the virtual human and another group of students (Group SP) interviewed a standardized patient, an actor trained to represent a medical condition. Standardized patient interviews are real-world interactions that allow students to role play the medical interview. They are used at medical schools all over the world to train and test medical students on interaction with patients. Standardized patient interviews are validated, effective, real-world simulations of patient interviews. They represent a gold standard to which to compare the virtual patient interaction.

4.1 Measures

4.1.1 Eliciting Critical Information

Participants were graded on their ability to elicit critical information from the patient. In an acute abdominal pain scenario, 12 critical pieces of information must be elicited to reach a correct diagnosis.

1. When did the pain start?
2. Where is the pain located?
3. What does the pain feel like?
4. Is the patient nauseous?
5. Has the patient vomited?
6. Does the patient have an appetite?
7. Has the patient had any unusual bowel movements?
8. Is the patient sexually active?
9. When was the patient's last period?
10. Has the patient had any unusual vaginal discharge?
11. Does the patient have a fever?
12. Has the patient had any unusual urinary symptoms?

Students that elicited seven of the 12 items received a passing grade. Each group was graded by two parties:

Group SP: The standardized patient graded participants by noting the critical information she revealed in the interview. Medical experts also graded the interactions.

Group VH: The virtual patient system graded students by logging the critical information she revealed in the interview. Medical experts also graded the interactions.

4.1.2 Interaction Behavior

Interactions were examined for behavioral differences between the two groups. Oviatt et al. observed that spontaneous disfluencies (false starts, hesitations, etc.) occur less in machine-human interaction than in human-human interaction [32]. Therefore, interactions were assessed on the conversation flow. Conversation flow was graded by counting the number of confirmatory words, like "ok" and "mmhmm," used in the interview. Such phrases are often used when a person understands what the other is saying and wants to continue with the next topic. The expert observers also noted qualitative differences in conversation flow.

The interactions were also analyzed for empathetic behavior. Empathizing with the patient is a key component of building rapport. Empathy lets the patient know the doctor understands her situation [33]. Empathetic behavior is also an indicator of the participant's emotional involvement in the interaction. Participants' empathetic actions (e.g., saying "I know it hurts," acknowledging the patient's fears, etc.) were tallied.

4.1.3 Perceptions of the Interaction

The Maastricht Assessment of the Simulated Patient (MaSP) [34] is a validated survey used to evaluate standardized patients. To gather perceptions of the virtual and real interactions, participants filled out a modified MaSP focusing on authenticity and behavior. Questions on the MaSP include whether the patient is challenging/testing, if the patient maintains appropriate eye contact, and if the simulated patient could be a real patient.

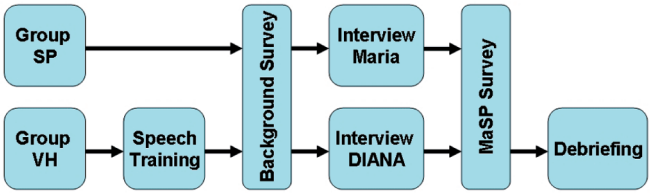

Fig. 3. Study procedure for groups SP and VH.

4.2 Participant Background

Group SP $(n = 8)$**:** Eight second-year medical students (four male and four female) from the University of Florida interviewed the standardized patient (SP). On average, this group interviewed sixteen SPs prior to this study.

Group VH $(n = 16)$**:** Nine medical (four first-years, one second-year, four third-years) and seven physician-assistant students (four first-years, two third-years, one fourth-year) from the Medical College of Georgia interviewed the virtual human. Seven were male, eight were female, and one did not specify a gender. On average, this group interviewed four SPs prior to this study.

4.3 Procedure

Fig. 3 summarizes the Study I procedure.

4.3.1 Pre-Experience

Participants arrived at a teaching and testing center where students routinely interview standardized patients. Each participant signed a consent form and filled out a background survey. They were then taken to an exam room and told their patient was inside. They were instructed to interview the patient, but not to do a physical exam.

Group SP: Participants put on a tracked hat for head gaze logging. The standardized patient also wore a tracked hat. The head gaze data will be analyzed at a later date.

Group VH: Participants put on a tracked, wireless microphone headset and a finger-worn infrared ring for gesture recognition. They also trained the system's speech recognition software to create a personalized voice profile.

4.3.2 Experience

The procedure for both groups mirrored the procedure students routinely follow when interviewing standardized patients. This allowed for a more valid comparison between the virtual and real interviews and also helped students feel more comfortable with the system.

Group SP: Before the experience, participants waited outside the medical exam room door. When the standardized patient was ready, the words "You may now start the station" were played from an overhead speaker. The students then entered the examination room. The standardized patient was inside, lying on an examination bed. The standardized patient was in character as soon as the participant entered the room, and the interview began immediately upon entering. Participants were given up to 10 minutes to conduct the interview. After eight minutes, a bell was played to warn participants that two minutes remained in the interview. After 10 minutes, the following words were played from the overhead speaker: "Time is up. Please leave the station." As the participants had all

Fig. 4. VIC (left) and DIANA (right) in the virtual exam room.

experienced standardized patient interviews at the testing facility before, they were familiar with this procedure and the audio cues. The audio cues were scheduled on a strict timer. It should be noted that the amount of time spent interviewing the standardized patient varied from participant to participant. Participants who finished early were allowed to leave the room and move on to the postexperience surveys.

Group VH: As in Group SP, participants waited outside the door of a medical exam room at the beginning of the experience. When the virtual human was ready, the study staff instructed participants to enter the room. Upon entering the room, participants sat in a chair and faced the projection of the virtual exam room. Fig. 4 shows the virtual scene presented by the system. The virtual instructor, VIC, stood in the background and the virtual patient, DIANA, lay on the examination bed in the foreground.

Participants were instructed to say "hello" to VIC to begin the interaction. VIC responded by guiding participants through a short tutorial on interacting with the system. After the tutorial, VIC told the participant she had 10 minutes to complete the interview. VIC then left the room so that the participant and DIANA could have privacy. The 10 minute timer began as soon as VIC left. At the eight-minute mark, VIC informed the participant that two minutes remained over the system speaker (without reentering the room). After 10 minutes, VIC returned to the room and ended the interaction. He then asked the participant for a diagnosis. After the participant stated their diagnosis, VIC thanked the participant and asked her to leave the room. As in Group SP, the amount of time spent interviewing the virtual human varied from participant to participant. Participants who finished early were allowed to leave the room and move on to the postexperience surveys.

One might be concerned that the presence of a virtual instructor could lead Group VH to believe they were being observed. Actually, it is standard practice for instructors to observe interactions with standardized patients via closed-circuit camera. This practice was followed for both groups, and all participants were aware that that they were being observed.

4.3.3 Postexperience

Group SP: Participants related their perceptions of the standardized patient by filling out the MaSP survey (see Section 4.1).

Group VH: Participants related their perceptions of the virtual human by filling out the MaSP survey. They were then debriefed to obtain qualitative feedback about the experience.

5 Study I: Results and Analysis

This section reports similarities and differences between the virtual and real interpersonal scenarios. Interactions were similar on the information elicited from the patient and participant behavior. However, closer analysis reveals rapport-building behavior (e.g., comforting the patient) was less sincere with the virtual human. This stemmed from the virtual human's limited expressiveness. The virtual human's vocal and facial expression of pain did not match the real human's expression of pain. Finally, participant perceptions of authenticity were conflicting. Some measures indicated the virtual and real interactions were similarly authentic, while others indicated the real interaction was more authentic.

5.1 Statistical Analysis and Nomenclature

Throughout this section, a two-tailed Student's T-Test is used to test for significant differences ($\alpha < 0.05$). Note that items where differences were not found are not guarantees of similarity. Instead, the term similarity in this article denotes an inability to show statistically significant differences and a reasonable closeness in the mean and standard deviation. Statistical equivalence tests are gaining acceptance [35], and we plan to use them in future work.

Throughout this section, statistics are presented of the form $M \pm S$, where M is a mean and S is a standard deviation. Unless otherwise noted, M_{SP} represents a fraction of the participants in Group SP, and M_{VP} has the same meaning for Group VP.

5.2 Content Measures

The content of the real and virtual interactions were similar. The virtual human and standardized patient were asked the same questions and they responded to the questions similarly. Furthermore, both groups tried to use empathy with their patient. These similarities indicate the virtual human interaction meets the content goals of the standardized patient experience.

5.2.1 Eliciting Critical Information

The purpose of the medical interview is to gather critical information needed to reach a diagnosis. Therefore, participants were graded on their ability to elicit critical information from the patient. Participants were graded by the patient (virtual or real) and the expert observers. Three observers scored Group SP and four observers scored Group VH on the critical information metric. Two of the observers were the same for both groups.

To assess the variability between the expert's observations, the total score on the critical information metric was correlated pair-wise across observers. The lowest Pearson

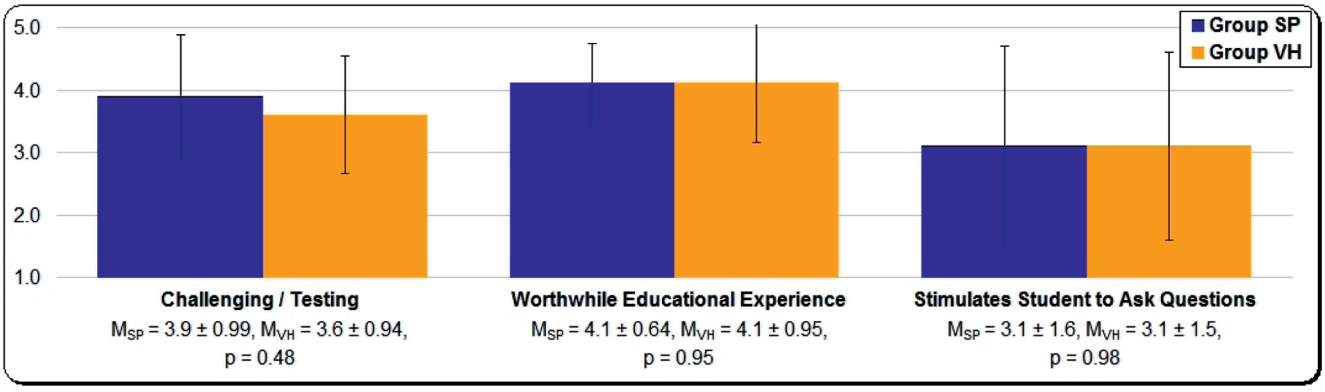

Fig. 5. Participants indicated (MaSP) that real and virtual interactions had similar educational value. $1 =$ Complete Disagreement, $5 =$ Complete Agreement.

correlation between the expert observers was $r = 0.86$. All correlations were significant at $p = 0.006$ or lower. This significant, large positive correlation indicates there was little interobserver variation on the critical information metric. Observer scores were combined by averaging due to the low interobserver variation. Mean values (M_{SP} and M_{VP}) represent the fraction of participants that elicited a critical piece of information.

According to the expert observers, participants asked the virtual human and standardized patient the same critical questions. As the virtual human's responses were based on the standardized patient's responses, the answers to the critical questions were also similar in both groups. No difference was found for both easily discussed information ("The pain is sharp and stabbing," $M_{SP} = 1 \pm 0$, $M_{VH} = 0.8 \pm 0.35$, $p = 0.12$) and sensitive information ("I am sexually active," $M_{SP} = 0.54 \pm 0.5$, $M_{VH} = 0.45 \pm 0.52$, $p = 0.72$). Final scores on eliciting the 12 critical items ($M_{SP} = 6.3 \pm 1.7$, $M_{VH} = 5.5 \pm 2.1$, $p = 0.37$) and the fraction of students who received a passing grade ($M_{SP} = 0.5 \pm 0.54$, $M_{VH} = 0.36 \pm 0.5$, $p = 0.58$) were also similar. This suggests that a virtual human can sufficiently perform the role of a real person in a constrained, information exchange scenario.

In normal standardized patient interactions, grades are usually given by the standardized patient instead of expert observers. Therefore, students were also graded by the patient (virtual or real) they spoke to. According to the patient grades, questions about the location/progression of the pain ($M_{SP} = 0.25 \pm 0.46$, $M_{VH} = 1 \pm 0$, $p = 1E-6$), the fact that the patient is nauseated ($M_{SP} = 0.88 \pm 0.35$, $M_{VH} = 0.25 \pm 0.44$, $p = 0.0023$), and the fact that the patient is sexually active ($M_{SP} = 0.88 \pm 0.35$, $M_{VH} = 0.44 \pm 0.51$, $p = 0.042$) were not asked with the same frequency in both interactions.

Although differences were found in the patient grades, we defer to the experts grades because they are more reliable and consistent. Standardized patient grading is not always reliable because standardized patients are human. Whether consciously or subconsciously, they take other subjective factors into account when grading. Also, standardized patients grade during short breaks in between interactions with medical students. They do not have much time to consider grades because another student is waiting outside for the next interaction. The medical experts, on the other hand, watched the interactions on video afterward. They had ample time to review the video and make sure their grades were accurate. There was also a high degree of consistency between the experts grades, lending more strength to their observations. The SP's grades, however, were not correlated to any of the experts (at best, $r = -0.041$, $p = 0.923$). The higher reliability and consistency of the experts grades indicates that the real and virtual interactions were similar in eliciting critical information.

It should be noted that the virtual human is the only truly objective grader. This is because the virtual human graded participants by logging whatever information she revealed to them. The virtual human cannot take into account other factors when grading. In contrast to the standardized patient's grades, the virtual human's grades tend to match the experts grades closely. For example, the standardized patient's grades on sexual history differed by 34 percent from the experts grades, but the virtual human's grades on sexual history differed from the experts by only 1 percent. The virtual human's ability to grade similarly to a panel of medical experts is a clear advantage over the subjective grading of standardized patients.

5.2.2 Educational Goals

The similarities on content show that interactions with a virtual human can meet the educational goals of interactions with a real human. Participants in the virtual interaction were able to practice asking a patient questions, a key aspect of the medical interview. Furthermore, participants rated the virtual and real interactions' educational value similarly (see Fig. 5). One student said, "I thought it was really interesting, it was challenging and it was good to refresh my memory on a lot of communication and interviewing skills." Another student noted that the system allows one to practice the process of interviewing a patient without feeling nervous: "It was a lot less pressure than a real person, even a standardized patient. In there with the virtual patient, I wasn't worried about looking natural and confident ... looking natural to the real patient. I was out there taking time trying to figure out what's wrong with the patient." The virtual scenario was a valuable educational experience.

5.2.3 Empathy

Empathy was used with the standardized patient and virtual human. The number of times Group SP and VH expressed empathy to the patient was similar ($M_{SP} = 2.2 \pm 1.4$, $M_{VH} = 1.3 \pm 1.1$, $p = 0.44$). Group VH used empathy when the virtual human expressed fear about her pain. A typical empathetic response with the virtual human was: "Don't worry. I am going to help you." Group SP used similar empathetic statements. They also touched the standardized patient's arm and used a softer tone of voice to comfort her.

Empathy encourages patients to share information. By expressing empathy, participants were working toward their task of eliciting critical information from the patient. Also, the use of empathy is a sign that participants tried to engage the virtual human emotionally. This is encouraging, considering the virtual human is not a real person.

5.3 Behavior

5.3.1 Empathy

The expert observers noted that both groups tried to build rapport with their patient through empathy, but Group VH's empathetic behavior appeared less genuine. Group SP typically spoke naturally and used a soft tone of voice. Some participants touched the standardized patient's leg or the exam bed and held it there for a moment.

On the other hand, Group VH used a more rehearsed, robotic empathy. They responded to the virtual human's cry for help, but their lack of emotional expression and monotone voice made these empathetic responses appear insincere. Of course, participants could not touch the virtual human as she occupies the virtual space beyond the projection on the wall. However, no participant even tried to touch the image of the virtual human on the wall. In debriefings, one participant from Group VH said: "I'm (normally) really engaging with my patients. Even though it was very real, it was very cold and artificial. I couldn't get very involved." This comment hints that the poor expressiveness of the virtual human led participants to adapt their conversation style.

It was also clear that some students in Group VH bothered to use empathy because they are required to in interviews with standardized patients. Their training (and fear of a bad grade) compelled them to use empathy. However, they did not have to appear sincere since the virtual human was not capable of evaluating and responding to sincerity (or the lack thereof).

From an evaluation and training standpoint, these students gamed the system. They knew they could behave improperly with the virtual human without being penalized for it. For this system to be effective, it must be able to detect when students game the system and respond appropriately. The virtual human should make a comment or change her mood to make it obvious to the participant that their lack of sincerity is improper. Making the virtual human more sensitive to improper behavior will encourage students to be more sincere.

5.3.2 Conversational Behavior

The behaviors people used to manage the conversation with the virtual human were very different than the behaviors used to manage the conversation with the standardized patient. These differences were a result of the limitations of the virtual human's conversational architecture.

Context-Dependent Questions: The virtual human could not respond to context-dependent questions. As a result, context-dependent questions were used initially but were quickly abandoned. For example, if the virtual human said "I ate a sandwich," a typical follow up question would be "When?" Participants quickly learned the virtual human did not remember context from question to question. Instead, they rephrased the question: "When did you eat the sandwich?"

Rapid-Fire Questions: The difficulty with context-dependent questions and other conversational idiosyncrasies led participants to ask the virtual human questions in a rapid-fire fashion. One student noted: "I was forced to use choppy sentences and direct questions." This resulted in many students robotically going through a mental checklist of questions. Sometimes they paused to think of the next question to ask. One student remarked on the patient's behavior during pauses: "When we pause for 3 seconds the patient sometimes will volunteer information, but with the system, when you're quiet, she's quiet." The system was essentially one directional in nature. It only responded when it was asked a question. This was unnatural because real conversations are two-way.

Conversation Flow: The flow of the conversation was also unusual. Normally, people use confirmatory phrases to regulate the conversation flow. Confirmatory phrases are short, one word acknowledgements ("Yeah," "uh-huh," etc.) or repetitions of what was just said. For example, the standardized patient might say, "My stomach hurts a lot." The participant's response would be "OK. Your stomach hurts. Can you show me where the pain is?" Confirmatory phrases were used throughout standardized patient interviews to confirm what the patient said and signal the start of another question. Far fewer confirmatory phrases were used with the virtual human ($M_{SP} = 20 \pm 4.7$, $M_{VH} = 3.5 \pm 4.1$, $p = 6E - 5$).

5.4 Expressiveness: Virtual Human versus Standardized Patient

5.4.1 Differences

The expert observers noted several differences between the virtual human's and standardized patient's expressive behavior. The standardized patient spoke very little because she was in too much pain to speak. Her voice was low in tone and volume and was somewhat raspy. She almost always had a look of extreme pain on her face. Her facial expressions varied with motion to indicate how painful it was to move. Head-nodding, eye contact, and timely responses contributed to the participant's sense (gathered from the MaSP) that the standardized patient was listening.

The virtual human was much less expressive. Her voice had a regular volume and tone. Her face did not convey the right level of pain. She occasionally shifted her body or moved her hands, but her facial expressions did not change accordingly. Besides looking at the participant, the virtual human used no other explicit behaviors to indicate listening. Occasional delays in speech recognition produced

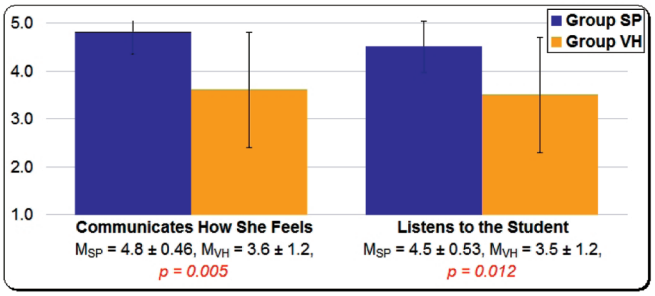

Fig. 6. Participants felt the virtual human was less expressive than the standardized patient (MaSP). $1 =$ Complete Disagreement, $5 =$ Complete Agreement.

delays in the virtual human's responses. Participants often interpreted this to mean the virtual human was not as engaged in the conversation. Feedback on the MaSP showed the standardized patient communicated how she felt better than the virtual human and appeared to be a better listener (see Fig. 6).

Previous results did not indicate the virtual human's lack of vocal expressiveness was a major deficiency of the system [2], and no significant difference was found between synthesized speech and more realistic recorded speech [36]. Therefore, no effort was put into improving DIANA's voice. However, this comparison was between different speech modes of the virtual system. When compared against the voice of the standardized patient, the lack of professional voice quality impacted results. This study shows that professional voice talent is necessary.

Differences in expressiveness were also pronounced because the virtual human animation tools made it difficult to create sophisticated expressive behaviors within a reasonable amount of time. The large difference in expressiveness suggests that effort must be invested to improve expressiveness before future studies are conducted.

5.4.2 Similarities

The only expressive behavior where the virtual human and standardized patient were similar was eye contact. Head tracking enabled the virtual human to look at the participant. This gaze behavior, life-size imagery, and rendering from the participant's perspective led Group VP to indicate on the MaSP that the virtual human used appropriate eye contact ($M_{SP} = 4 \pm 1.6$, $M_{VH} = 3.7 \pm 0.99$, $p = 0.61$). One Group VP participant said: "I felt that it was neat that they were life-size, you know, and that the patient is looking at you and talking to you."

5.4.3 Effects of Expressiveness

We hypothesize that the differences in patient expressiveness affected several measures. Global measures of authenticity (see Section 5.5) and impressions of the interaction overall ($M_{SP} = 9.5 \pm 0.53$, $M_{VH} = 6.6 \pm 2.0$ on a scale of 1 to 10, $p = 1E-4$) were lower in Group VH. Furthermore, although participants asked the virtual human and standardized patient similar questions, some behavior (see Section 5.3) was different with the virtual human. Group VH asked questions in a more direct, rapid-fire fashion, and changes to conversation flow were observed. Empathy was expressed, but the empathy was not as sincere as that seen in the real scenario. Participants suggested that the virtual human be more expressive: "I would suggest to have more emotions into them. Maybe if there was more feelings, more emotional expression." The effectiveness of virtual humans are strongly impacted by their expressiveness.

5.5 Authenticity

As standardized patients are simulations of real patients, it is useful to compare standardized patients to real patients to judge their authenticity. The MaSP (see Section 4.1.3) is a standardized survey filled out by medical students to assess standardized patient authenticity from a local and global perspective. Global measures look at overall impressions of the interaction, like whether the standardized patient acts like a real patient. Local measures look at specific components of the interaction, like whether the patient expressed pain realistically. Participants filled out the MaSP after interacting with the virtual human and standardized patient.

Global (big-picture) measures indicated the virtual interaction was less authentic than the real interaction. The virtual human appeared less authentic ($M_{SP} = 5 \pm 0.0$, $M_{VH} = 3.8 \pm 0.58$, $p = 9E-6$) and was less less likely to be considered a real patient ($M_{SP} = 4.8 \pm 0.46$, $M_{VH} = 3.8 \pm 1.1$, $p = 0.008$). Also, the virtual encounter was less similar to other standardized patient encounters ($M_{SP} = 4.5 \pm 1.1$, $M_{VH} = 2.5 \pm 0.94$, $p = 2.00E-4$).

However, local (subcomponent) measures mostly indicate the virtual and real scenarios were not different on authenticity. No differences were found on whether the patient simulated physical complaints unrealistically ($M_{SP} = 1.8 \pm 1.4$, $M_{VH} = 2.6 \pm 1.0$, $p = 0.096$), whether the patient answered questions in a natural manner ($M_{SP} = 2 \pm 1.4$, $M_{VH} = 2.9 \pm 1.2$, $p = 0.13$), and whether the patient appeared to withhold information unnecessarily ($M_{SP} = 4.1 \pm 1.2$, $M_{VH} = 3.4 \pm 1.2$, $p = 0.23$). A single difference was found on whether the patient's appearance fits the role ($M_{SP} = 5 \pm 0.0$, $M_{VH} = 4.3 \pm 0.47$, $p = 4.0E-04$).

Given the differences in behavior and expressiveness, it is surprising that the virtual and real interactions were considered similar on local authenticity measures. One would expect *any* real interaction to always be considered more authentic than its virtual counterpart. We hypothesize that the real and virtual interactions were similarly authentic on local measures because participants applied different standards when rating local authenticity. Upon examining debriefing comments, it became clear that Group VH evaluated the "humanness" of the virtual human, whereas Group SP judged the accuracy of the standardized patient to a real patient. This is similar to Usoh et al.'s conclusion that people apply different standards to real and virtual environments on presence questionnaires [30].

5.6 Poststudy Reflections

Study I's results should be considered preliminary because of the following study characteristics:

Sample Size: The population size, particularly in the case of the SP experience ($n = 8$), was too limited.

Participant Experience: For scheduling reasons, it was difficult to recruit medical students of equal experience

levels. This difference affected the study results. Students with less experience do not yet know what a good or bad interview with an SP is. They apply a different grading standard than the more experienced students. More experienced students likely conduct better interviews.

Different Institutions: Logistical issues made it difficult to recruit Group VH and SP from the same institution. Students from the Medical College of Georgia and the University of Florida have a tendency (and/or have been trained) to ask different questions.

Volunteer Participants: All participants were recruited volunteers. Volunteers are typically highly motivated students and probably do not represent an accurate sample of the general medical student population. They perform better than the average medical student, and they rate the virtual experience more positively. An analysis of the effect of volunteer participants will be published in a future article. Future studies should use nonvolunteers to ensure accurate sampling of the general medical student population.

VH Voice Fidelity: As part of a separate study, some Group VH participants spoke to a virtual human with a computer-generated voice, and others spoke to a virtual human with prerecorded real speech. No significant difference was found between the text-to-speech and real-speech conditions on all measures [36]. Therefore, the two groups were combined together in our analysis. However, as virtual human expressiveness has been identified as affecting results, recorded professional talent should always be used.

Inexperience with the Virtual Human: The virtual human interaction was a new experience for Group VH, but the standardized patient experience was familiar to Group SP. This experience gap between groups is a potential confounding factor. Group VH's behavior may have been different because they needed more time to become comfortable with the system.

That said, we believe that students' experience with the real interaction partially transfers to the virtual and decreases any confounding effects. This transfer occurs because of the various ways the virtual interaction mimics the real interaction. Students experience the virtual interaction in the same medical exam rooms as in the real interaction. The projected virtual exam room is modeled after the real room. It has the same color walls, the same dimensions, the same kind of patient bed and so on. The virtual human mimics real-world symptoms of appendicitis. Even the study procedure was modeled after the participants' normal experiences with standardized patients.

Study I's results also provide evidence that experience transfers from the real to the virtual interaction. The fact that students asked the same questions in both real and virtual interactions implies that Group VH brought their experiences with them into the virtual interview. Furthermore, in a paper currently under review, a strong correlation was found in interaction skills between SP and VH interviews. Students who do well in VH interactions also do well in SP interactions. Likewise, students who do poorly in VH interactions do poorly in SP interactions. This could not be possible unless experience transfers from the real to virtual.

5.7 Summary

Study I shows that an information gathering scenario can be simulated effectively with life-size, interactive virtual humans. However, more complex scenarios that incorporate communication skills, like rapport-building, are more difficult to simulate because participants behave differently with virtual humans. We hypothesize that, as the expressiveness of virtual humans are improved, behavior with virtual humans will become more similar to that used in real interpersonal interactions. Study I also showed that interaction authenticity is difficult to measure with subjective surveys. Only objective measures should be used to assess interaction authenticity.

6 Study II: Motivation and Design

Study II was conducted four months after Study I. The study's goals were to 1) gain more insight into how behavior is different with virtual humans, 2) strengthen Study I's results with a larger sample ($n = 58$), and 3) address potential confounds in Study I. Study II differed from Study I in that:

- New behavioral measures were added to characterize rapport with the virtual human and standardized patient.
- All participants were second-year medical students from the University of Florida. Restricting participants to the same institution and year of study controlled for differences in training and skill level.
- The virtual human used more expressive, prerecorded speech. This guaranteed that the voice fidelity of the virtual human was the same for all participants.
- The system was integrated into a patient communication course. The course gave access to a larger sample of the general medical student population. Instead of using volunteer participants, medical students were randomly selected to interact with the virtual human as part of their coursework. Participation was not compulsory.

As in Study I, students were split into two groups. Group VH ($n_{VH} = 33$) interviewed the virtual human and Group SP ($n_{SP} = 25$) interviewed a standardized patient.

6.1 Measures

A panel of five medical experts graded the interactions on content and rapport-building behavior. Participant perceptions of the interaction were not gathered in Study II.

6.1.1 Content

As in Study I, expert observers noted if participants asked the patient about twelve critical pieces of information needed to reach a diagnosis (see Section 4.1).

Reaching a diagnosis may be easier with a detailed picture of a patient's medical history. Therefore, participants were graded on whether they elicited information on five history categories: social history, family history, history of present illness, medical history, and review of systems. Patient history information is not critical to reaching a diagnosis, but it can be helpful in narrowing down the list of topics to ask about.

6.1.2 Rapport-Building Behavior

Differences in rapport-building behavior in Study I motivated a more detailed analysis in Study II. Expert observers graded participants on the following expanded behavioral measures:

Process and Etiquette: In addition to gathering facts, a medical student should follow several guidelines related to the interview process and etiquette. Students should introduce themselves to the patient, use transitional statements to progress through the interview, and conduct the interview in an orderly fashion. Such guidelines help doctors collect information logically and communicate clearly with the patient.

Empathy: Sincere empathy is key to building rapport with patients. Experts noted whether empathy was used in the interviews and how spontaneous it was. They also characterized participants on the following 7-point Likert scales: Good/Bad, Strong/Weak, Active/Passive, Valuable/Worthless, Powerful/Powerless, Fast/Slow, Talkative/Quiet, Helpful/Unhelpful, and Deep/Shallow. Together, these scales provide a descriptive breakdown of empathetic behavior.

Nonverbal Communication: People use nonverbal behavior, often subconsciously, to communicate attitudes and feelings. For example, sustained eye contact, forward body lean and proper head nodding communicate attentiveness and an overall positive attitude [37]. Experts graded participants on nonverbal communication because it contributes significantly to rapport with the patient.

6.2 Procedure

Study I's procedure was modified to match the patient communication course procedure. Normally, students arrive at a medical education facility at a predetermined hour. Each student is assigned two medical exam rooms where standardized patients are waiting. As soon as overhead speakers play "You may now start the station," the students simultaneously enter their first assigned room. The participants interact with the standardized patient for up to 10 minutes. At the eight-minute mark, a warning bell is played to inform students that two minutes remain. When students complete their interview, they use any remaining time, plus a two minute break, to summarize what they learned about the patient and suggest a course of action. After the two minute break, the students repeat the process with their second standardized patient. Once a student is finished with the second interview, she is free to go.

As our study was integrated into the course, we followed this procedure precisely, with only slight deviations. Appointments were made so that Group VH could do speech training before the interactions. During the study, Group VH participants were assigned one room with a virtual human with symptoms of appendicitis and one room with a standardized patient with different symptoms. Likewise, Group SP participants were assigned one room with a standardized patient with symptoms of appendicitis and one room with a standardized patient experiencing different symptoms.

Due to the strict course schedule, and a desire to make the virtual interaction like the real one, VIC was removed from the virtual interaction. As a result, the start, two-minute warning, and end sounds were all played from the facility's speakers. Also, participants were not given a system tutorial. Initially, there was some concern that the loss of VIC's tutorial would lead to participant confusion. In fact, removing the tutorial made the experience more familiar to students. Students were able to start the interview immediately as they usually do in standardized patient interactions.

Following the course schedule made it difficult to collect data from students. There was no time for participants to fill out the MaSP survey used in Study I, nor were participants debriefed for comments. On the other hand, an advantage of following the course schedule was that study participation was a more familiar experience to students.

7 Study II: Results and Analysis

Study II strengthened our findings in Study I. The content of the virtual and real interactions remained similar, and rapport again was lower in the virtual interaction. Process and etiquette were followed less strictly, and nonverbal behavior conveyed less interest and a less positive attitude toward the virtual human. As in Study I, the limited expressiveness of the virtual human was a factor in changing participant behavior.

7.1 Interobserver Reliability

To assess the relative agreement of the expert observers, three summary scores were correlated across observers. The first summary score was a tally of the objective, "yes/no" measures—critical information, patient history information, and process and etiquette. The second summary score was an average of the subjective empathy descriptors. The final summary score was an average of the nonverbal communication measures (eye contact, body lean, etc.).

Table 1 shows the pair-wise Pearson correlation of the observers on the three summary measures. Most observers are reasonably correlated with each other $(r > 0.4)$ and have a less than 5 percent $(\alpha < 0.05)$ chance of being correlated due to chance. This implies the observers rated the interactions similarly.

On all measures, Observer 05 was not highly correlated with at least one observer and/or the correlations were more likely to come from chance than from true agreement. Therefore, observer O5 was culled from the study. Observer O3's nonverbal behavior observations were also culled from the study for the same reasons. After culling these observations, the ratings were combined into a single measurement by averaging.

7.2 Content Measures

7.2.1 Eliciting Critical Information

As in Study I, participants elicited the same critical information from the virtual human and standardized patient. Both groups were equally likely to elicit 11 of 12 critical pieces of information from the patient. Fig. 7 compares the overall performance of eliciting critical information in Study I and Study II. Not only was the overall performance similar across groups, it was also similar across studies. The consistency of eliciting critical

TABLE 1
The Pearson Correlation between Observers

Pearson Correlation (r)

	Observer	O1	O2	O3	O4	O5
Yes/No Objective	O1	1	0.795	0.654	0.822	X
	O2	0.795	1	0.862	0.597	X
	O3	0.654	0.862	1	0.600	0.754
	O4	0.822	0.597	0.600	1	0.166
	O5	X	X	0.754	0.166	1
Empathy	O1	1	0.82	0.584	0.839	X
	O2	0.82	1	0.484	0.738	X
	O3	0.584	0.484	1	0.472	0.911
	O4	0.839	0.738	0.472	1	0.467
	O5	X	X	0.911	0.467	1
Nonverbal Communication	O1	1	0.424	0.755	0.83	X
	O2	0.424	1	0.307	0.454	X
	O3	0.755	0.307	1	0.670	0.592
	O4	0.83	0.454	0.670	1	0.990
	O5	X	X	0.592	0.990	1

Significance (p)

	Observer	O1	O2	O3	O4	O5
Yes/No Objective	O1	-	9E-6	3E-5	2E-5	X
	O2	9E-6	-	1E-10	0.040	X
	O3	3E-5	1E-10	-	0.002	0.083
	O4	2E-5	0.040	0.002	-	0.754
	O5	X	X	0.083	0.754	-
Empathy	O1	-	3E-6	3E-4	7E-6	X
	O2	3E-6	-	0.004	0.006	X
	O3	3E-4	0.004	-	0.017	0.012
	O4	7E-6	0.006	0.017	-	0.351
	O5	X	X	0.012	0.351	-
Nonverbal Communication	O1	-	0.049	2E-7	1E-5	X
	O2	0.049	-	0.083	0.138	X
	O3	2E-7	0.083	-	2E-4	0.215
	O4	1E-5	0.138	2E-4	-	2E-4
	O5	X	X	0.215	2E-4	-

An X indicates no correlation due to insufficient observer overlap.

information across studies strengthens the assertion that content was similar in the real and virtual interactions.

A single difference was found on whether the student elicited the location of the pain ($M_{SP} = 0.75 \pm 0.36$, $M_{VH} = 0.91 \pm 0.16$, $p = 0.02$). System logs show the virtual human often revealed the pain location even when not directly asked about it. This was due to errors in matching noisy input speech to responses in the virtual human's database. As part of our future work, the system's response matching thresholds will be tuned to reduce false positives. This will help prevent the system from revealing information that has not been asked for.

7.2.2 Patient History

As an additional content measure, experts graded participants on their ability to elicit patient history information. Patient history provides more background that can help the student reach a diagnosis. Overall, participants elicited less patient history information from the virtual human. For example, Group VH was less likely to gather family medical history ($M_{SP} = 0.5 \pm 0.48$, $M_{VH} = 0.22 \pm 0.38$, $p = 0.018$) and the history of the present illness ($M_{SP} = 0.85 \pm 0.26$, $M_{VH} = 0.63 \pm 0.32$, $p = 0.008$).

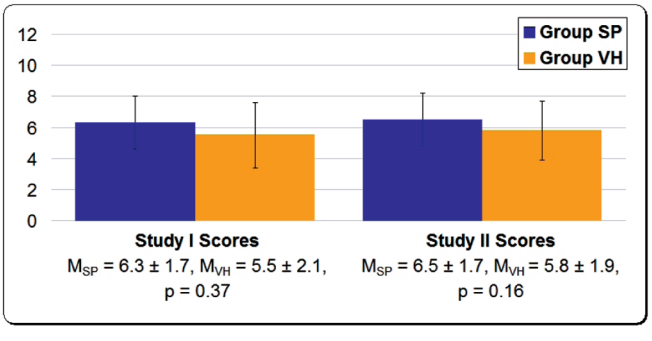

Fig. 7. Eliciting information was similar across both groups and both studies.

Differences in gathering patient history highlight how people adapt their behavior to the limitations of the virtual human. A medical student would normally use multiple follow-up questions to explore these topics. Follow-up questions are difficult for the virtual human to handle because they require knowledge of context. As in Study I, participants discovered that they cannot ask the virtual human context-dependent questions, and they adapted their behavior appropriately. As part of our future work, we plan on tracking context over the course of the interview. This will allow the virtual human to determine that follow up questions refer to previous questions.

It should be noted that no differences were found on whether participants asked about social history ($M_{SP} = 0.43 \pm 0.41$, $M_{VH} = 0.33 \pm 0.37$, $p = 0.33$). This is likely because social history questions (e.g., "Do you drink alchohol?") have very few follow-up questions. Also, participants may have avoided social history because it is a sensitive subject. Approximately 60 percent of participants did not ask social history questions.

Differences in patient history do not necessarily indicate that the content of the interaction was different overall. Gathering critical information is a much more important part of the interview than gathering patient history and should be weighted more strongly in the overall assessment of the interview content. Despite differences in gathering patient history, the overall content of the virtual and real interviews was similar.

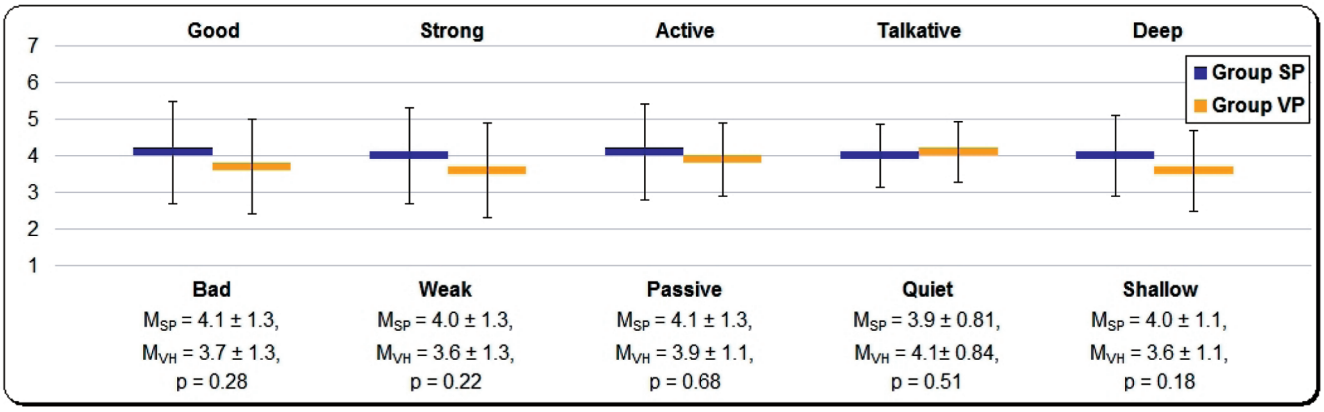

Fig. 8. Expert ratings of empathy on descriptive scales were too subjective to yield any differences.

7.3 Rapport-Building Behavior

As in Study I, participant behavior led to the impression that rapport was lower with the virtual human. Empathetic behavior was used, but was not sincere. Furthermore, some process and etiquette guidelines were not followed. Finally, nonverbal behavior communicated less interest and a poorer attitude toward the virtual human.

7.3.1 Process and Etiquette

Medical students are taught to follow specific process and etiquette guidelines in the medical interview. These guidelines make it easier to collect patient information logically and help build rapport with the patient.

Most process and etiquette guidelines were followed in the virtual and real interactions. Participants introduced themselves ($M_{SP} = 0.793 \pm 0.36$, $M_{VH} = 0.68 \pm 0.39$, $p = 0.28$), explored their patient's concerns ($M_{SP} = 0.87 \pm 0.30$, $M_{VH} = 0.86 \pm 0.18$, $p = 0.96$), and ended the interview appropriately ($M_{SP} = 0.54 \pm 0.43$, $M_{VH} = 0.58 \pm 0.34$, $p = 0.69$). These results are surprising because the virtual human does not "care" whether these guidelines are followed. The virtual human does not act differently if participants end the interview inappropriately. Clearly, participants applied rules from the real-world to this virtual interpersonal interaction.

When process and etiquette were abandoned, it was because the virtual human could not handle them properly. For example, Group VH conducted the interview in a less logical and orderly fashion ($M_{SP} = 0.87 \pm 0.25$, $M_{VH} = 0.53 \pm 0.35$, $p = 0.0001$). Participants did not have a logical orderly conversation with the virtual human because the virtual human is incapable of having a conversation in a logical order. For example, a student may be discussing headaches with the virtual human. If speech recognition misinterprets the next question to be about fever, the virtual human will suddenly change the topic and respond about her fever. This unexpected topic change shows participants that the system does not care about the order of questions. Therefore, participants do not bother interacting with the system in any logical order.

As mentioned previously, future versions of the system will address this by tracking the current topic, or context, of the conversation. This will allow the virtual human to determine when a query changes the topic and if the change in topic is logical. If the topic change is unexpected, the virtual human can ask the user to repeat the question to confirm.

7.3.2 Empathy

As in Study I, participants in both groups responded empathetically to their patient ($M_{SP} = 0.79 \pm 0.29$, $M_{VH} = 0.69 \pm 0.42$, $p = 0.34$). However, the sincerity of the empathy was lower with the virtual human. Group VH's empathetic behavior remained robotic and disengaged in Study II. Twenty-seven percent of Group VH used spontaneous empathy versus 84 percent in Group SP ($p = 4.38E - 6$). A close to significant difference was found on the overall quality of the empathy on a scale of 1 to 4 ($M_{SP} = 2.7 \pm 0.85$, $M_{VH} = 2.3 \pm 0.84$, $p = 0.08$).

Surprisingly, specific, descriptive ratings of empathetic behavior in the virtual and real interactions were not different. For example, Fig. 8 shows that both groups were rated similarly on descriptive scales like "good/bad," "weak/strong," and "active/passive." This is in stark contrast to the overall sense that empathy behavior was poorer with the virtual human. We hypothesize that no differences were found on these scales because they are too subjective. The expert raters could not objectively rate abstract concepts like "weak/strong." As part of our future work, we are exploring the use of objective behavioral measures to augment these subjective measures.

7.3.3 Nonverbal Communication

Nonverbal communication is critical to rapport-building because it communicates a variety of feelings and attitudes. Group VH's nonverbal behavior expressed lower rapport with the virtual human (see Fig. 9). They used less forward body lean and nodded less. These behaviors were inappropriate because they are associated with lower interest and a poorer attitude ([37]). Not surprisingly, Group VH appeared less attentive and had a less positive attitude with the virtual human (see Fig. 9).

It should be noted that expert ratings of participants' eye contact were similar with the virtual and real human. This was also seen in Study I, where participants indicated the virtual human and standardized patient maintained good eye contact. The virtual human constantly looked at the participant throughout the interview, influencing participants to reciprocate and maintain eye contact. Future studies should incorporate an eye tracker to confirm this result and determine the amount of eye contact more accurately.

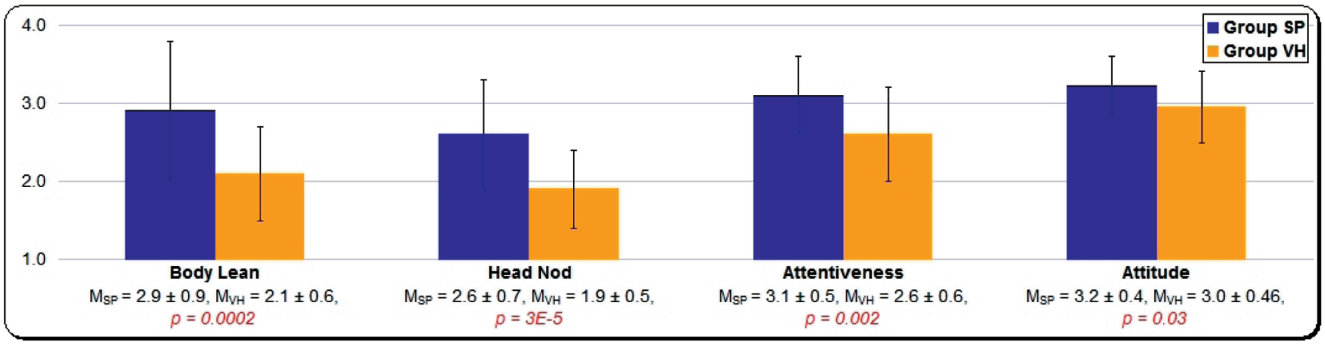

Fig. 9. Body lean and head nod behavior communicated a poorer attitude and less attentiveness to the virtual human. $1 = \text{Very Poor}$, $4 = \text{Very Good}$.

7.4 Expressiveness of the Virtual Human

Study I hypothesized that the virtual human's lower level of expressiveness (compared to the standardized patient) played a role in changing rapport with the virtual human. One might expect then that the virtual human's expressive, prerecorded voice in Study II would lead to fewer differences on rapport-building behavior. On the contrary, differences in behavior remained.

Clearly, expressiveness must be improved further. The virtual human should use everyday conversational idiosyncracies, like stopping to think and saying "um" and "uh." Her face should convey more pain. Her body should be less rigid, yet still enough to convey the pain that moving creates. Her responses to queries should be immediate. This list is only a small sample of the expressive abilities that must be improved.

7.5 Summary

Study II strengthens the findings of Study I with a larger sample $(n = 58)$ and fewer potential confounding factors. Content measures remained similar while behavior related to rapport showed strong differences between the virtual and real interactions. Differences on nonverbal communication provide more evidence that rapport-building is lower with the virtual human. As in Study I, these differences likely stemmed from the virtual human's limited expressiveness.

8 Conclusions

Using the medical interview as a platform, two studies were conducted that compare a virtual interpersonal interaction to an effective, standardized, real-world interaction. Expert observations and participant feedback indicated that the virtual human was less expressive than the standardized patient. This led to less rapport-building with the virtual human. However, the virtual interaction was similar to the real interaction on content measures. Participants gathered the same information from the virtual human and standardized patient.

The studies also show that interaction authenticity and participant empathy cannot be assessed easily. Global measures showed that the real scenario was more authentic, but local measures suggest—on a component level—that the virtual scenario was similar to the real scenario. A similar contradiction was seen in expert ratings of empathy. While participants appeared to use insincere empathy with the virtual human, subjective, descriptive ratings of empathy found no differences. These results lead to the following guidelines for constructing and evaluating effective interpersonal simulators:

Construction: An interpersonal scenario where information gathering is the main task can be effectively simulated if the content of the interaction is similar to its real-world counterpart. However, scenarios that incorporate communication skills, like rapport-building, are more difficult to simulate because virtual humans do not meet the standard of expressiveness set by real humans. The expressiveness of virtual humans must be improved to elicit natural behavior from users.

Evaluation: Evaluating an interpersonal simulator is difficult to do objectively. More objective, physically-based measures of authenticity and behavior must be developed.

9 Future Work

Comparing a virtual interpersonal simulation to an effective, standardized real-world counterpart is a key step in learning how to build effective interpersonal simulators. Now that this step has been taken, we can start exploring the effect of several variables on the effectiveness of interpersonal simulators. Most important is the effect of varying the virtual human's expressive behavior. Other important variables to study include the virtual human's mesh quality, rendering quality, and the system display device (projector, monitor, and HMD).

To support these studies, we have rebuilt the Virtual Patient system. The new system supports higher-quality virtual human models and animations. We are now able to create animations using the same tools video game and movie effects artists use. Programmatic control over these animations allows systematic control over the breadth, depth, and quality of the virtual human's expressive behaviors. Improvements have also been made to the natural language system to improve the conversation flow with the virtual human. The virtual human responds to input faster and tracks the context of the interview.

Most critical to understanding why participant behavior changes with virtual humans is the development of objective, physical measures of behavior. Sensors, like the microphone and reflective markers users already wear, will be used to characterize physical behavior. The following subset of behaviors that impact perceived rapport with the patient will be tracked:

- Posture—Does the student adopt open, friendly postures?
- Gaze—Does the student look at the patient, or elsewhere?
- Facial Expressions—Does the student use appropriate, friendly, reassuring facial expressions?
- Speaking Time—Does the student talk too much or too little? Are there long pauses while the student thinks?

To help with interpreting this behavioral data, a tool for visualizing interactions between real and virtual humans is being developed. Visualization will provide a focusing lens through which we can analyze the collected behavioral data.

ACKNOWLEDGMENTS

The authors thank the Harrell Professional Development and Assessment Center at the University of Florida and the Medical College of Georgia for the use of their facilities. They thank Maria Martinez for being their standardized patient and the voice of DIANA in Study II. They thank Emily Gorovsky for help with editing this paper and being the voice of DIANA in Study I. Last, they acknowledge the University of Florida Alumni Graduate Fellowships for partially funding this work.

REFERENCES

[1] K. Johnsen, R. Dickerson, A. Raij, C. Harrison, B. Lok, A. Stevens, and D.S. Lind, "Evolving an Immersive Medical Communication Skills Trainer," *Presence: Teleoperators and Virtual Environments,* vol. 15, no. 1, pp. 33-46, Feb. 2006.

[2] K. Johnsen, R. Dickerson, A. Raij, B. Lok, J. Jackson, M. Shin, J. Hernandez, A. Stevens, and D.S. Lind, "Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills," *Proc. IEEE Conf. Virtual Reality,* 2005.

[3] R. Dickerson, K. Johnsen, A. Raij, B. Lok, J. Hernandez, and A. Stevens, "Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction," *Proc. Int'l Conf. Human-Computer Interface Advances for Modeling and Simulation,* 2005.

[4] A. Raij, K. Johnsen, R. Dickerson, B. Lok, M. Cohen, A. Stevens, T. Bernard, C. Oxendine, P. Wagner, and D.S. Lind, "Interpersonal Scenarios: Virtual ≈ Real," *Proc. IEEE Conf. Virtual Reality,* pp. 59-66, 2006.

[5] N.I. Badler, C.B. Phillips, and B.L. Webber, *Simulating Humans: Computer Graphics, Animation, and Control.* Oxford Univ. Press, 1993.

[6] N.E. Alessi and M.P. Huang, "Evolution of the Virtual Human: From Term to Potential Application in Psychiatry," *CyberPsychology & Behavior,* vol. 3, no. 3, pp. 321-326, 2000.

[7] K.R. Thórisson and J. Cassell, "Why Put an Agent in a Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue," *Proc. Conf. Lifelike Computer Characters,* 1996.

[8] *Embodied Conversational Agents,* J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, eds. MIT Press, 2000.

[9] V. Vinayagamoorthy, A. Steed, and M. Slater, "Building Characters: Lessons Drawn from Virtual Environments," *Proc. Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop,* pp. 119-126, 2005.

[10] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater, "Building Expression into Virtual Characters," *Proc. Eurographics State of the Art Reports,* 2006.

[11] C. Nass and Y. Moon, "Machines and Mindlessness: Social Responses to Computers," *J. Social Issues,* vol. 56, no. 1, pp. 81-103, 2000.

[12] D.-P. Pertaub, M. Slater, and C. Barker, "An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience," *Presence: Teleoperators and Virtual Environments,* vol. 11, no. 1, pp. 68-78, 2002.

[13] M. Garau, M. Slater, S. Bee, and M.A. Sasse, "The Impact of Eye Gaze on Communication Using Humanoid Avatars," *Proc. SIGCHI Conf. Human Factors in Computing Systems,* 2001.

[14] J.N. Bailenson, E. Aharoni, A.C. Beall, R.E. Guadagno, A. Dimov, and J. Blascovich, "Comparing Behavioral and Self-Report Measures of Embodied Agents Social Presence in Immersive Virtual Environments," *Proc. Seventh Ann. Int'l Workshop PRESENCE,* 2004.

[15] J.N. Bailenson, J. Blascovich, A.C. Beall, and J.M. Loomis, "Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments," *PRESENCE: Teleoperators and Virtual Environments,* vol. 10, no. 6, pp. 583-598, 2001.

[16] C. Zanbaka, P. Goolkasian, and L.F. Hodges, "Can a Virtual Cat Persuade You? The Role of Gender and Realism in Speaker Persuasiveness," *Proc. Conf. Human Factors in Computing Systems,* pp. 1153-1162, 2006.

[17] C. Zanbaka, A. Ulinski, P. Goolkasian, and L.F. Hodges, "Effects of Virtual Human Presence on Task Performance," *Proc. Int'l Conf. Artificial Reality and Telexistence,* pp. 174-181, 2004.

[18] K.R. Thórisson, "Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People," *Proc. First ACM Int'l Conf. Autonomous Agents,* pp. 536-537, 1997.

[19] R.W. Hill, Jr., J. Gratch, S. Marsella, J. Rickel, W. Swartout, and D. Traum, "Virtual Humans in the Mission Rehearsal Exercise System," *Kynstliche Intelligenz (KI J.),* vol. 17, no. 4, 2003.

[20] A. Manganas, M. Tsiknakis, E. Leisch, M. Ponder, T. Molet, B. Herbelin, N. Magnenat-Thalmann, D. Thalmann, M. Fato, and A. Schenone, "The Just VR Tool: An Innovative Approach to Training Personnel for Emergency Situations Using Virtual Reality Techniques," *J. Information Technology in Healthcare,* vol. 2, no. 6, pp. 399-412, 2004.

[21] S. Balcisoy, R. Torre, M. Ponder, P. Fua, and D. Thalmann, "Augmented Reality for Real and Virtual Humans," *Computer Graphics Int'l,* pp. 303-308, 2000.

[22] N.I. Badler, C.A. Erignac, and Y. Liu, "Virtual Humans for Validating Maintenance Procedures," *Comm. ACM,* vol. 45, no. 7, pp. 56-63, 2002.

[23] A.A. Rizzo, D. Klimchuk, R. Mitura, T. Bowerly, J.G. Buckwalter, K. Kerns, K. Randall, R. Adams, P. Finn, I. Tarnanas, C. Sirbu, T.H. Ollendick, and S.-C. Yeh, "A Virtual Reality Scenario for All Seasons: The Virtual Classroom," *Proc. 11th Int'l Conf. Human-Computer Interaction,* July 2005.

[24] P. Emmelkamp, M. Krijn, A. Hulsbosch, S. de Vries, M. Schuemie, and C. van der Mast, "Virtual Reality Treatment versus Exposure in Vivo: A Comparative Evaluation in Acrophobia," *Behaviour Research and Therapy,* vol. 40, pp. 509-516, 2002.

[25] B.O. Rothbaum, L. Hodges, S. Smith, J.H. Lee, and L. Price, "A Controlled Study of Virtual Reality Exposure Therapy for the Fear of Flying," *J. Consulting and Clinical Psychology,* vol. 68, no. 6, pp. 1020-1026, 2000.

[26] M. Billger, "Colour Appearance in Virtual Reality: A Comparison Between a Full-Scale Room and a Virtual Reality Simulation," *Proc. Ninth Congress Int'l Colour Assoc.,* 2001.

[27] D. Wuillemin, G. van Doorn, B. Richardson, and M. Symmons, "Haptic and Visual Size Judgements in Virtual and Real Environments," *Proc. First Joint Eurohaptics Conf. and Symp. Haptic Interfaces for Virtual Environment and Teleoperator Systems,* pp. 86-89, 2005.

[28] I. Heldal, R. Schroeder, A. Steed, A.S. Axelsson, M. Spant, and J. Widestrom, "Immersiveness and Symmetry in Copresent Scenarios," *Proc. IEEE Conf. Virtual Reality,* pp. 171-178, 2005.

[29] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder, "Small-Group Behavior in a Virtual and Real Environment: A Comparative Study," *PRESENCE: Teleoperators and Virtual Environments,* vol. 9, no. 1, pp. 37-51, 2000.

[30] M. Usoh, E. Catena, S. Arman, and M. Slater, "Using Presence Questionnaires in Reality," *PRESENCE: Teleoperators and Virtual Environments,* vol. 9, no. 5, pp. 497-503, 2000.

[31] F. Lang, R. McCord, L. Harvill, and D.S. Anderson, "Communication Assessment Using the Common Ground Instrument: Psychometric Properties," *Family Medicine,* vol. 36, no. 3, pp. 189-198, 2004.

[32] S.L. Oviatt, P.R. Cohen, M. Wang, and J. Gaston, "A Simulation-Based Research Strategy for Designing Complex NL Systems," *Proc. ARPA Human Language Technology Workshop,* 1993.

[33] J.L. Coulehan and M.R. Block, *The Medical Interview: Mastering Skills for Clinical Practice,* third ed. F.A. Davis Co, 1997.

[34] L.A. Wind, J. van Dalen, A.M.M. Muijtjens, and J.-J. Rethans, "Assessing Simulated Patients in an Educational Setting: The MASP (Maastricht Assessment of Simulated Patients)," *Medical Education,* vol. 38, no. 1, pp. 39-44, 2004.
[35] D.F. Parkhurst, "Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation," *BioScience,* vol. 51, no. 12, pp. 1051-1057, 2001.
[36] R. Dickerson, K. Johnsen, A. Raij, B. Lok, T. Bernard, A. Stevens, and D.S. Lind, "Virtual Patients: Assessment of Synthesized versus Recorded Speech," *Proc. Conf. Medicine Meets Virtual Reality 14,* 2006.
[37] A. Mehrabian, *Nonverbal Communication.* Aldine-Atherton, 1972.

Andrew B. Raij graduated from Northwestern University with a BS degree (2001) and from the University of North Carolina at Chapel Hill with an MS degree (2004) in computer science. He is a PhD student in the Computer and Information Science and Engineering Department at the University of Florida. His research interests include computer graphics, virtual environments, virtual humans, visualization, and projector-camera systems. For more information, see http://www.cise.ufl.edu/~raij.

Kyle Johnsen is a PhD student at the University of Florida. His research areas include virtual humans and human computer interaction.

Robert F. Dickerson graduated from the University of Florida with a BS (2006) degree in computer engineering. He is a PhD student at the University of Virginia. His research interests include virtual environments and scalable graphics rendering. He is a student member of the IEEE. For more information, see http://www.cs.virginia.edu/~rfd7a.

Benjamin C. Lok graduated from the University of Tulsa with a BS degree (1993) in computer science, and from the University of North Carolina at Chapel Hill with an MS (1997) and PhD degrees (2002). He is an assistant professor in the Computer and Information Science and Engineering Department at the University of Florida. He is an adjunct assistant professor in the Department of Surgical Oncology at the Medical College of Georgia. His research areas include computer graphics, virtual environments, and human-computer interaction. He is a member of the IEEE. For more information, see http://www.cise.ufl.edu/~lok.

Marc S. Cohen received his medical degree from the University of Miami School of Medicine. He completed his surgical residency at Boston University and a urological residency and research fellowship at the University of Texas Medical Branch. He is currently a professor and Residency Program Director in the Department of Urology at the University of Florida, College of Medicine. Dr. Cohen has a special interest in clinician patient communication skills. He is a faculty member of the Institute for Health Care Communication and speaks nationally and internationally on health care communication.

Margaret Duerson, PhD, is an associate professor of medicine at the University of Florida and the director of the Harrell Professional Development and Assessment Center.

Rebecca Rainer Pauly, MD, FACP, is an associate professor of medicine, chief of the Division of Internal Medicine, and serves as Associate Chair of Medicine for Medical Student Education at the University of Florida. She is the course director of Essentials of Patient Care I-IV, a two-year continuum focused on communication, history, physical exam skills, and medical decision-making. Dr. Pauly's research focuses on medical education topics such as advanced communication, physical exam, and the medical interview.

Amy O. Stevens is the division chief of the OB/GYN Department at the NF/SG VA Health System and holds a courtesy assistant professor appointment with the Department of Surgery at the University of Florida. She is board certified in Internal Medicine, having completed her medical (1995) and internal medicine training (1998) at the University of Florida and an OB/GYN residency (2005) at the Medical College of Georgia. Her research interests include patient safety and simulation.

Peggy Wagner received the PhD degree in social psychology from the University of Florida. She is a professor in the School of Medicine at the Medical College of Georgia. She is the Director of Research and Faculty Development in the Department of Family Medicine, Medical Educational Research in the School of Medicine and the Clinical Skills Program. Her areas of interest are physician-patient communication, behavioral medicine, cultural competency, and patient education.

D. Scott Lind, MD, is the Jarrell Distinguished Professor and chief of the Section of Surgical Oncology within the Department of Surgery at the Medical College of Georgia. Dr. Lind has a strong interest in medical education and a scientific approach to the study of educational problems. His educational research focus is on the development and evaluation of new technologies, particularly simulation, in medical education. He is part of a collaboration involving the Medical College of Georgia and the University of Florida to develop and evaluate virtual patients for teaching and testing clinical skills.