Interpersonal Scenarios: Virtual ≈ Real?

Andrew Raij, Kyle Johnsen, Robert Dickerson, Benjamin Lok[1]

Department of Computer and Information Science and Engineering, University of Florida

Marc Cohen, Amy Stevens[2]

SHANDS VA Hospitals, University of Florida

Thomas Bernard, Christopher Oxendine, Peggy Wagner, D. Scott Lind[3]

Medical College of Georgia

ABSTRACT

This paper reports on a study to examine the similarities and differences in experiencing an interpersonal scenario with real and virtual humans. A system that allows medical students to interview a life-size virtual patient using natural speech and gestures was used as a platform for this comparison. Study participants interviewed either a virtual patient or a standardized patient, an actor trained to represent a medical condition. Subtle yet substantial differences were found in the participants' rapport with the patient and the flow of the conversation. The virtual patient's limited expressiveness was a significant source of these differences. However, overall task performance was similar, as were perceptions of the educational value of the interaction.

CR Categories and Subject Descriptors: J.3. Life and Medical Sciences, K.3 Computers and Education, H.5 Information Interfaces and Presentation

Additional Keywords: Virtual Characters, Multimodal Interaction, Human-Computer Interaction, Medical Education, Immersive Virtual Environments

1 INTRODUCTION

Advances in rendering, audio, and animation allow virtual humans to be presented with increasing levels of fidelity. Improvements in tracking, gesture recognition, and voice recognition also enable natural means of interaction. This combination of high-fidelity output and natural input has led to research into the use of virtual humans as partners in interpersonal scenarios.

The concept of interpersonal, virtual humans raises an important question: How is experiencing an interpersonal scenario with a virtual character similar to – and different from – experiencing one with a real person? Clearly there must be differences, as no one would be "fooled" by a virtual character into thinking they were interacting with a real person. But, in which ways can they be similar? What are the key differences? This study explores these questions by analyzing self-reported and behavioral measures of participants in similar interpersonal interactions with real and virtual people. The study asks:

- Are post-encounter impressions similar?
- Are empathy and other emotions and attitudes similarly expressed?
- Which social constructs are followed?

These questions must be explored to:

1. Determine the extent to which interpersonal scenarios can be simulated with virtual humans
2. Identify how component technologies need to improve to enable effective interpersonal virtual human systems

As current technology does not allow simulation of any general interpersonal scenario, this work employed a specific standardized and constrained scenario. In [13], a system was presented that

Figure 1 – A real interpersonal interaction (right) and an equivalent virtual interpersonal interaction (left). In the real interaction, a medical student interviews a real standardized patient. In the virtual interaction, the medical student interviews a virtual patient.

[1]{raij, kjohnsen, rfd, lok}@cise.ufl.edu
[2]{cohenms, steveao}@surgery.ufl.edu
[3]{tbernard, coxendine}@students.mcg.edu
{pwagner, dlind}@mail.mcg.edu

allowed medical students to interview a life-size virtual patient using natural speech and gestures. The highly-constrained structure of the doctor-patient interview and natural interaction with the virtual humans led participants to (anecdotally) treat the virtual patient similarly to a standardized patient [9].

This system was used as a platform to explore the similarities and differences between real and virtual interpersonal scenarios. Twenty-four medical students conducted a patient-doctor interview:

They were divided into two groups:

- Group VP interacted with a virtual patient representing a specific condition (Figure 1 - left).
- Group SP interacted with a standardized patient, an actor representing the same condition (Figure 1 - right)

The interactions were compared using the following metrics:

- Student assessment of patient (virtual or real) performance
- Expert assessment of student performance
- Patient (virtual and real) assessment of student performance
- Participant behavior during the interaction

Several key similarities between interacting with real and virtual humans were found. Participants in both groups performed equally well on overall task performance (condition diagnosis), and impressions of the educational value of both interactions were similar. Also, measures of the components of the virtual patient's authenticity were similar to that of the standardized patient.

Subtle, yet substantial differences were found in the flow of the conversation and participant rapport with the patient. The virtual patient's limited expressiveness was a significant source of these differences.

These results show that the primary goals of the interaction were met even though the virtual human is easily distinguishable from a real human. However, significant effort should be applied towards creating expressive virtual humans to improve the effectiveness of interpersonal scenarios.

2 Previous Work

2.1 Effective Virtual Humans

Several virtual character systems have been developed for training, teaching, and education. Thórisson [21] presented an interactive guide named Gandalf that takes users on tours of the solar system. USC's Institute for Creative Technologies has created virtual experiences to train military personnel in interpersonal leadership [12]. The Just VR system [15] allows a medical trainee to interact with a virtual assistant to assess and treat a virtual victim. The Human Modeling and Simulation Group at the University of Pennsylvania uses virtual humans for task analysis and assembly validation [2].

Researchers have worked to establish the basis of effective virtual humans. Badler et al [3] suggest that virtual humans "should move or respond like a human" and "must exist, work, act and react within a 3D virtual environment." Alessi and Huang [1] expand these rules further in the context of virtual character applications for psychology. They highlight the need for virtual humans to be social, emotionally expressive, and interactive. Virtual humans should be able "to 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." Thorisson and Cassell [22] agree that emotional expression is likely important, but non-verbal behaviors that support the conversation, e.g. hand gestures for pointing at objects being discussed and looking at the user to indicate attention, are more significant. In a review of virtual character research, Vinayagamoorthy et al [24] concluded that 1) the behavioral and visual fidelity of virtual humans must be consistent, and 2) a virtual character's expressions should be appropriate for the context of the application. Nass has pioneered research into the affective power of computers and intelligent agents. Their work has shown that people can ascribe very human characteristics to computers, such as helpfulness, usability, and friendliness [16].

2.2 Human Behavior with Virtual Humans

There is growing evidence that people treat virtual people similarly to real people. Pertaub et al [18] noted participants with a fear of public speaking talking to an audience of virtual humans reported similar anxieties as when speaking to an audience of real people. Garau et al [11] showed that realistic, task-appropriate avatar eye-gaze behavior led to improved communication between the people represented by the avatars.

Bailenson et al have shown that people manage their personal space when interacting with virtual humans similarly to when they interact with real humans. They found that people displayed a tendency to put more space between them and an embodied tutor than they did with strangers [4], and participants maintained more distance from embodied agents than inanimate virtual objects [5]. Female participants maintained more distance from embodied agents that maintained eye contact than with agents that did not.

2.3 Virtual Versus Real Experiences

We have found little work that directly compares real and simulated interpersonal scenarios. However, researchers have compared other virtual environments to their real counterparts. In the psychology domain, Emmelkamp et al [10] compared the reactions of acrophobes in similar virtual and real environments. Using standardized measures of acrophobia, the authors found that exposure therapy in the virtual environment was as effective as therapy in the real environment. Rothbaum et al [19] compared virtual and real exposure therapy for those with fear of flying. Results show experiencing a virtual airplane is just as effective as experiencing a real plane in reducing fear of flying anxiety. Both types of therapy are significantly better than no therapy at all.

Others have looked at human perception of real and virtual stimuli. To explore the use of VR for lighting and color planning in buildings, Billger [6] examined the perception of color in virtual and real environments. Wuillemin et al [26] looked at differences in 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.

Slater et al [20] looked at the social behavior of small groups in real and virtual environments. Immersed participants were viewed as leaders by their peers in the virtual scenario but not in the real environment. Furthermore, group accord was higher in the real environment.

3 Virtual Patient System

3.1 Doctor-Patient Interview Overview

The medical practitioner's goal during the interview is to get the key facts of the patient's condition and come to a differential diagnosis, a list of possible conditions the patient may have. Also important is communicating with the patient and addressing their fears (rapport). Research has shown medical communication skills can be taught and practiced and do not just improve over time with clinical experience [14]. The interaction between patients and doctors during initial diagnosis is an interpersonal scenario 1) which can be simulated despite the technology

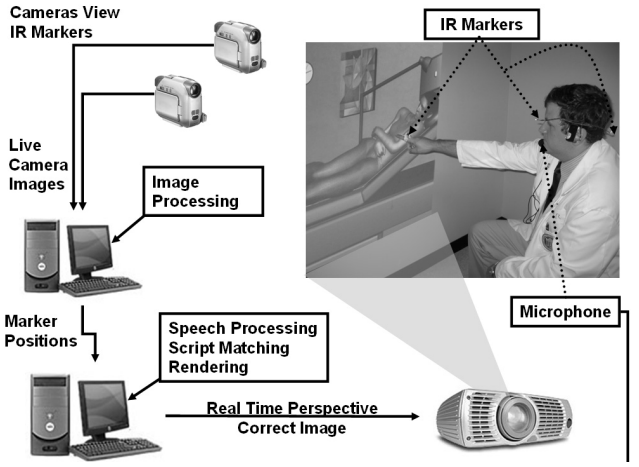

Figure 2 – System overview

limitations, 2) where practice is costly, and 3) where immersion and fidelity is important.

3.2 System Description

The system used in the virtual patient interaction was similar to the one described in [13] (Figure 2). On the wall of a real examination room, a virtual exam room (modeled as an extension of the real room) was projected at life-size. Participants interacted with the virtual patient, DIANA (a 5'6", 18-year-old Caucasian female), and a virtual instructor, VIC, with speech and gestures. Commercial speech recognition software and a simple algorithm for parsing utterances [9] allowed the participant to talk to VIC and DIANA naturally within the scope of the scenario. An index finger was tracked, allowing the participant to localize DIANA's pain with simple pointing gestures. The participant's head was tracked to render the scene from her perspective, and DIANA and VIC can maintain eye contact with the participant.

DIANA's appearance and responses are based on a real standardized patient, Maria, trained in acute abdominal pain (AAP). AAP is a common ailment and a basic scenario in patient-doctor interaction and communication skills education. The goal was to create similar real and virtual interpersonal interactions with an AAP patient.

4 Study Design

A preliminary study was designed to explore similarities and differences between the real and virtual interpersonal scenarios. One group of students (Group VP) at the Medical College of Georgia (MCG) interviewed the virtual patient, and another group of students (Group SP) at the University of Florida (UF) interviewed a real standardized patient. Standardized patient interviews are real interpersonal interactions that represent a gold standard to which to compare the virtual patient interaction.

4.1 Measures

The following measures were used to compare the virtual and real interactions.

4.1.1 Student Assessment of Standardized/Virtual Patient

The Maastricht Assessment of the Simulated Patient (MaSP) [25] is a validated survey used to evaluate standardized patients. A modified MaSP, focusing on patient authenticity and behavior, was used to assess the standardized and virtual patient. Questions on the MaSP include whether the patient is challenging/testing, whether the patient maintains appropriate eye contact and whether the simulated patient could be a real patient. Participants in both groups filled out the MaSP after the interview.

4.1.2 Standardized/Virtual Patient Assessment of Student

A grading system jointly devised at MCG and UF was used to assess student performance in the standardized and virtual patient interviews. In an acute abdominal pain scenario, twelve critical pieces of information must be elicited from the patient to reach a correct differential diagnosis. These include when the pain started, the location of the pain, and whether the patient is sexually active. Eliciting seven of the twelve items constitutes a passing grade.

Group SP: The standardized patient graded participants by noting what critical information she revealed in the interview.

Group VP: The virtual patient system graded students by logging the critical information it revealed in the interview.

4.1.3 Expert Assessment of Student

Medical experts from both institutions watched video tapes of Group VP and Group SP. The interactions were graded using the critical information measure discussed in the previous section.

4.1.4 Behavioral Measures

Interactions were examined for behavioral differences between the two groups. Oviatt observed spontaneous disfluencies (false starts, hesitations, filled pauses, repairs, fragments) occur less frequently in machine-human interaction than in human-human interaction [17]. Therefore, interactions were assessed on qualitative and quantitative measures of the flow of the conversation. The number of confirmatory words, like "ok" and "mmhmm," were counted. Such phrases are often used when a person understands what the other is saying and wants to continue with the next topic. Obvious qualitative differences were also noted as they were observed.

The interactions were also analyzed for empathetic behavior. Empathizing with the patient is an important skill that lets the patient know the doctor understands her situation [7]. Empathetic behavior is also an indicator of the participant's emotional involvement in the interaction. The number of empathetic actions (e.g. saying "I know it hurts," acknowledging the patient's fears or touching the patient) that a participant performed was recorded.

4.2 Participant Background

Group SP (*n*=8) - Eight 2nd-year medical students (four male and four female) from UF interviewed the real standardized patient (Maria). On average, this group had interviewed sixteen standardized patients prior to this study.

Group VP (*n*=16) - Nine medical students (four 1st-years, one 2nd-year, four 3rd-years) and seven physician-assistant students (four 1st-years, two 3rd-years, one 4th-year) from MCG interviewed the virtual patient, DIANA. Seven were male, eight were female, and one did not specify a gender. On average, this group had interviewed four standardized patients prior to this study.

4.3 Procedure

Figure 3 summarizes the study procedure for Groups SP and VP.

4.3.1 Pre-Experience

Participants in Group VP and SP 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. Participants 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.

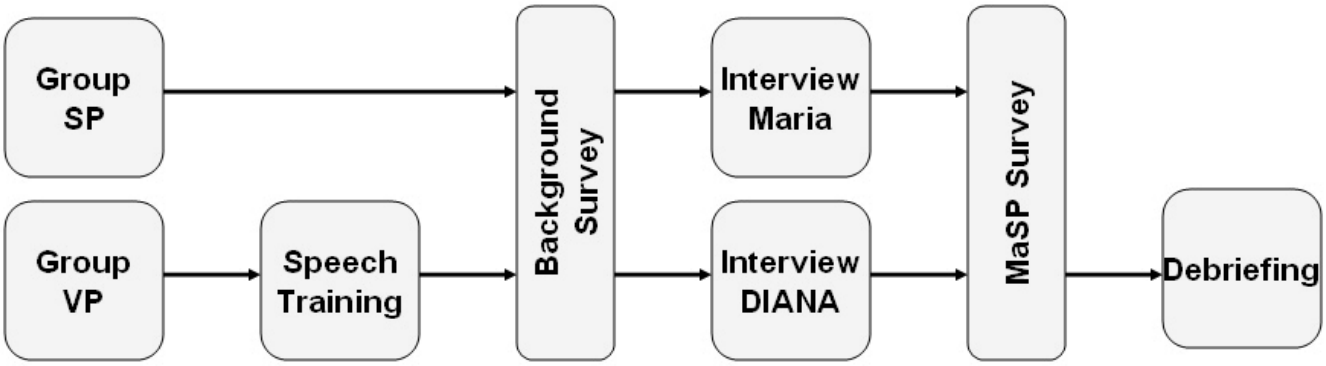

Figure 3 – Study procedure for groups SP and VP

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

Group VP – Participants wore 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

Group SP – Upon entering the room, participants saw the real standardized patient lying on an examination room table. Participants were given up to ten minutes to interview the patient. After eight minutes, participants were told two minutes remained in the interview. Participants left the room when time was up.

Group VP – Upon entering the room, participants sat in a chair and faced the projection of the virtual exam room. Figure 4 shows the virtual scene presented by the system. The virtual instructor, VIC, stood in the background and the virtual patient, DIANA, lay on an examination table in the foreground. VIC began the experience with a tutorial explaining how to interact with the system. VIC then told the participant to begin the interview and left the room so that the participant and DIANA could have privacy. VIC returned and alerted the participant when two minutes remained. After ten minutes, VIC ended the session and asked the participant for her differential diagnosis. VIC thanked the participant and asked her to leave the room.

4.3.3 Post-Experience

Group SP – Participants assessed the standardized patient by filling out the MaSP survey (Section 4.1.1). The standardized patient also graded the participant on the critical information checklist for the acute abdominal pain scenario (Section 4.1.2).

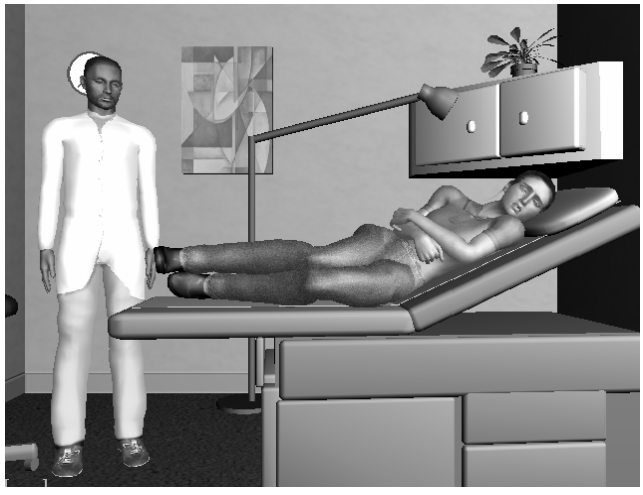

Figure 4 – VIC (left) and DIANA (right) in the virtual exam room

Group VP – Participants assessed the virtual patient by also filling out the MaSP survey. This group was then debriefed to obtain qualitative feedback about the experience. As the virtual patient system tracked the information it revealed in the interview, it automatically logged this data to grade participant performance.

Experts – Medical experts from both institutions independently watched video recordings of the real and virtual interactions. They assessed Group SP and VP on the critical information metric.

5 Results and Analysis

This section reports similarities and differences between the virtual and real interpersonal scenarios in five areas: participant performance, participant behavior, scenario authenticity, patient expressiveness, and educational goals. M_{SP} and M_{VP} are means for the standardized and virtual patient, respectively.

A two-tailed Student's T-Test is used to test for significant differences ($\alpha < 0.05$). Equivalence tests were also performed but did not yield any significance due to the small sample size. Note that items where differences were not found are not guarantees of ***equivalence***. Instead, the term ***similarity*** in this article denotes the inability to show differences between the VP and SP conditions; additional studies (VP n = 66, SP n = 33) have been conducted to allow for stronger equivalence comparisons. These results are currently being analyzed.

5.1 Participant Performance

Participant performance was measured in terms of the twelve critical items a student must elicit from the patient to reach a diagnosis (see Section 4.1.2). Mean values presented here represent the fraction of participants that elicited the critical item.

Overall performance was not different between the two groups, and both groups tended to elicit the same information from the patient. From a task performance standpoint, participants had similar interactions with both the virtual and real patient.

Table 1 - Expert assessment of performance on 12 critical items the student must elicit from the patient. Overall performance and two sample items (in quotes) are shown. Students asked the same questions and performed similarly in both interactions.

Critical Information	M_{SP}	M_{VP}	α
"The pain is sharp and stabbing"	1.0 ± 0	0.8 ± 0.35	0.12
"I am sexually active"	0.54 ± 0.5	0.45 ± 0.52	0.72
Final Score out of 12	6.3 ± 1.7	5.5 ± 2.1	0.37
Passed (≥ 7 items elicited)	0.5 ± 0.54	0.36 ± 0.5	0.58

5.1.1 Similarities

Participants tended to ask the VP and SP the same questions. Experts graded Group VP and SP on the critical information metric. The expert assessment shows that both groups asked all critical items a similar number of times (see Table 1). No difference was found for both easily discussed information ("The pain is sharp and stabbing") and more sensitive information ("I am sexually active"). Final scores and the number of students who received a passing grade (≥ 7 items elicited) were also similar between the two groups.

The VP and SP also graded the participants on the critical information metric. Except for three items (discussed in the next section), the grades again show both groups asked the same questions and performed equally well on the task of getting the key facts of the patient's illness. The medical expert reviewers agreed that at a high level, the interactions and task performance of Group VP and Group SP were similar.

5.1.2 Differences

A comparison of grades given by the SP and VP show differences on some critical items. Significant differences were found on three of the twelve critical items: the location/progression of the pain ($M_{SP} = 0.25 \pm 0.46$, $M_{VP} = 1 \pm 0$, $\alpha = 1E\text{-}6$), the fact that the patient is nauseated ($M_{SP} = 0.88 \pm 0.35$, $M_{VP} = 0.25 \pm 0.44$, $\alpha = 0.0023$) and the fact that the patient is sexually active ($M_{SP} = 0.88 \pm 0.35$, $M_{VP} = 0.44 \pm 0.51$, $\alpha = 0.042$). These are discussed in more detail in Section 5.4.

5.1.3 Analysis

Participants had the same conversation with both the virtual and standardized patient. The similarities in the information elicited show that both groups asked about the same questions and heard the same responses from the patient. **This result suggests that a virtual character can sufficiently perform the role of a real person in a constrained, information exchange scenario.**

Furthermore, **task performance was similar in both interactions.** Both groups received similar grades and a similar percentage of participants received a passing grade on the experience. This supports the **external validation of the virtual scenario as having a strong correlation to its real world counterpart.** It also shows participants put the same effort into achieving the goals of a virtual interpersonal interaction as they would in a real one.

While the real and virtual patients graded scenarios similarly to each other, there is an inconsistency with the grades given by the experts on three critical items. Performance was likely different for these three items. However, the remaining nine items were elicited a similar number of times, and participants received a similar percentage of passing grades in both scenarios. Task performance was similar between the two groups no matter which set of graders were used for comparisons.

5.2 Behavioral Measures

Similar conversational behaviors were seen in both groups. However, there were subtle differences in the participants' rapport with the patient and the flow of the conversation.

5.2.1 Similarities

Both groups expressed empathy to the patient. The number of times Group SP and VP expressed empathy to the patient was similar ($M_{SP} = 2.2 \pm 1.4$, $M_{VP} = 1.3 \pm 1.1$, $\alpha = 0.44$). Medical practitioners use empathy to build rapport with patients. Empathy helps strengthen the patient-doctor bond and can encourage the patient to open up and tell the doctor more about her situation. Empathetic behavior includes saying "I know it hurts," acknowledging the patient's fears, and touching the patient.

5.2.2 Differences

Although the number of empathetic responses was similar between groups, Group VP's empathy appeared less genuine. Group SP typically spoke naturally and used a soft tone of voice when expressing empathy. Some participants touched the SP's leg or the exam bed and held it there for a moment. Although Group VP was also empathetic to their patient, their empathetic responses tended to be more rehearsed. Also, the physical wall between the virtual and real exam room made it impossible for participants to touch the virtual patient.

Group VP's interactions were less natural than Group SP's. Group VP asked questions in a direct, rapid-fire fashion. One student noted: "I was forced to use choppy sentences and direct questions." Many appeared to robotically go through a mental checklist of questions. Sometimes they paused for a while 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."

Context-dependent questions were used initially but were abandoned because the system did not respond properly. For example, if the VP said "I ate a sandwich," then a typical follow up question in Group SP would be "What time?" Group VP quickly learned that the system did not remember context from question to question. Instead, they asked "What time did you eat the sandwich?"

Group SP used more confirmatory phrases than Group VP. Confirmatory phrases regulate the flow of the conversation. They can be as short as one word acknowledgements ("Yeah", "uh-huh," "ok," etc.) or complete repetitions of what was said. For instance, the SP 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 the SP interview to confirm what the SP said and signal the start of another question. A very significant difference between the groups was found on the number of confirmatory phrases used ($M_{SP} = 20 \pm 4.7$, $M_{VP} = 3.5 \pm 4.1$, $\alpha = 6E\text{-}5$).

5.2.3 Analysis

Group VP adapted their conversational style to the limitations of the virtual patient. They asked questions in a more constrained manner and appeared to be less engaged. They did not ask context-dependent questions nor did they use confirmatory phrases to control the flow of the conversation. In a real interview, a patient would be bothered and confused by this style of conversation. They might even complain about it. Participants considered this style acceptable for the virtual patient interview because they knew the patient would not be bothered by it.

Participants tried to use empathy to build rapport with the virtual patient. Building rapport is important because it encourages patients to share information. By expressing empathy, participants were working towards their task of eliciting critical information from the patient and reaching a diagnosis. Also, the use of empathy is a sign that participants tried to engage the patient emotionally. This is encouraging, considering the virtual patient is not a real person.

Group VP did not express empathy as sincerely as Group SP. In debriefings, one participant from Group VP 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." The comment hints at the poor expressiveness of the virtual character, discussed in detail in Section 5.4. It is possible that by improving the expressiveness of the virtual patient, participants will be more likely to express sincere empathy.

5.3 Interaction Authenticity

This section examines student impressions of patient authenticity. Means are on a Likert scale from 1-5. While differences were found in direct (big-picture) measures of authenticity, no differences were found on indirect (subcomponent) measures. Analysis shows a battery of indirect measures will yield a clearer picture of authenticity than direct measures.

Table 2 – Direct (big picture)* and indirect (subcomponent) measures of patient authenticity.

Question	M_{SP}	M_{VP}	α
The SP simulates physical complaints unrealistically.	1.8 ± 1.4	2.6 ± 1.0	0.096
The SP answers questions in a natural manner.	2 ± 1.4	2.9 ± 1.2	0.13
The SP appears to withhold information unnecessarily.	4.1 ± 1.2	3.4 ± 1.2	0.23
This encounter is similar to other SP encounters that I've experienced.*	4.5 ± 1.1	2.5 ± 0.94	2E-4
The SP might be a real patient.*	4.8 ± 0.46	3.8 ± 1.1	0.008

5.3.1 Similarities

The virtual and real scenarios were equivalent on indirect (subcomponent) measures of authenticity. Participant responses were similar when asked to rate whether the patient simulates physical complaints unrealistically, whether the patient answers questions in a natural manner, and whether the patient appears to withhold information unnecessarily (see Table 2).

5.3.2 Differences

Differences were found on direct (big-picture) measures of authenticity. Participant responses show differences in whether the patient appears authentic (M_{SP} = 5 ± 0.0, M_{VP} = 3.8 ± 0.58, α = *9E-6*), whether the encounter is similar to other standardized patient encounters they've experienced and whether the patient might be a real patient. One difference was found on an indirect measure of authenticity: whether the patient's appearance fits the role (M_{SP} = 5 ± 0.0, M_{VP} =4.3 ± 0.47, α = *4E-4*).

5.3.3 Analysis

Indirect measures are better than direct measures at assessing the authenticity of a virtual interpersonal scenario. The inconsistencies on authenticity show that Group VP and SP applied different standards to big-picture questions, such as overall realism. Upon examining debriefing comments, it became clear Group VP evaluated the 'humanness' of the virtual patient, whereas Group SP judged the accuracy of the standardized patient to a real patient. This result is similar to Usoh et al's conclusion that people apply different standards when assessing real and virtual environments on presence questionnaires [23]. The indirect measures focused attention on individual aspects of the interaction. This allowed participants to specifically assess components, as opposed to deriving their own interpretations of overloaded terms such as "realism" and "natural".

Furthermore, **indirect measures make explicit what features of the system need to be improved.** For example, the significant difference in "whether DIANA's appearance fits the role" makes it clear that DIANA's appearance needs to be improved. Direct measures of authenticity should never be used. A battery of indirect measures will yield a clearer picture of virtual patient authenticity.

5.4 Virtual Character Expressiveness

The virtual patient was not nearly as expressive as the standardized patient. This affected the flow of the conversation and the amount of information the student was able to elicit from the patient (see Section 5.1).

Table 3 – The virtual patient was not nearly as expressive as the standardized patient.

Question	M_{SP}	M_{VP}	α
The SP maintained appropriate eye contact for this scenario.	4 ± 1.6	3.7 ± 0.99	0.61
The SP communicates how she/he felt during the session.	4.8 ± 0.46	3.6 ± 1.2	0.005
I can judge from the reactions of the SP whether he/she listens to the student.	4.5 ± 0.53	3.5 ± 1.2	0.012

5.4.1 Similarities

Participants felt the virtual and standardized patients used equally appropriate eye contact. The virtual patient was programmed to look at the participant. This gaze behavior, life-size imagery, and rendering the exam room from the participant's perspective contributed to the sense that the virtual patient used appropriate eye contact (see Table 3). 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."

5.4.2 Differences

The standardized patient expressed herself very differently from the virtual patient. Feedback showed the SP communicated how she felt better than the VP and appeared to be a better listener.

Behavioral analyses of the standardized and virtual patient highlight these differences. The standardized patient said very little during the interview 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. The SP's expressions varied to indicate how painful it was to move. Head-nodding, eye contact, and timely responses contributed to the sense that the SP was listening.

The virtual patient was much less expressive. Her voice had a regular volume and tone. The expression on her face did not convey enough pain. She occasionally shifted her body or moved her hands, but her facial expressions did not change accordingly. Besides looking at the participant during the interview, the virtual patient used no other explicit behaviors to indicate listening. Occasional, slight delays in speech recognition produced delays in the VP's responses. This was often interpreted as the VP was not as engaged in the conversation.

Previous results did not indicate the virtual patient's lack of vocal expressiveness was a major deficiency of the system [13], and no significant difference was found between synthesized speech and more realistic recorded speech [8]. 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 a standardized patient, the lack of professional voice quality impacted results. This study shows that the expressiveness of the patient's voice must be improved.

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

5.4.3 Analysis

The expressiveness of real people sets the bar very high for virtual characters. Clearly, the system failed to meet the standard of expressiveness set by the standardized patient. This affected more than just measures of expressiveness. Direct measures of authenticity and impressions of the interaction overall ($M_{SP} = 9.5 \pm 0.53$, $M_{VP} = 6.6 \pm 2.0$ on a scale of 1 to 10, $\alpha = 1E\text{-}4$) were significantly lower. Furthermore, although participants generally asked the VP and SP the same questions (Section 5.1), **the way participants talked to the VP was affected by expressiveness.** Participants asked questions in a more direct, rapid-fire fashion, and changes to conversational flow were observed. Empathy was expressed, but the empathy was not as sincere as that seen in the real scenario (see section 5.2). These behaviors are not appropriate for real interpersonal scenarios. Participants specifically suggested that the VP be more expressive: "I would suggest to have more emotions into them. Maybe if there was more feelings, more emotional expression."

Differences in performance may also be a result of the virtual patient's poor expressive behavior. Some data suggests Group SP tended to ask certain questions more than Group VP (see Section 5.1). It is possible the virtual patient's expressive behavior did not trigger the same questions from participants that the standardized patient did. For example, the VP did not look overly pale nor did she look like she wanted to vomit. As a result, participants may not have thought to ask if the VP was nauseated.

Clearly, virtual characters in interpersonal scenarios cannot be truly effective without being expressive. More research and effort is required to make virtual characters expressive.

5.5 Educational Goals

Not only was task performance similar between both scenarios (see Section 5.1), but participants felt the real and virtual scenario had similar educational value.

Table 4 – The educational value of both interactions were similar.

Question	M_{SP}	M_{VP}	α
The SP is challenging / testing the student.	3.9 ± 0.99	3.6 ± 0.94	0.48
I found this a worthwhile educational learning experience.	4.1 ± 0.64	4.1 ± 0.95	0.96
The SP stimulates the student to ask questions.	3.1 ± 1.6	3.1 ± 1.5	0.98
I would use this as a practice tool.	4.9 ± 0.35	4 ± 1.2	0.019

5.5.1 Similarities

The virtual and real scenarios were similar in student impressions of the educational value of the experience (see Table 4). Both groups rated the experiences similarly on whether the patient is challenging/testing the student, whether the interaction is a worthwhile educational learning experience and whether the patient stimulates the student to ask questions.

5.5.2 Differences

A difference was found in whether students would use the virtual patient as a practice tool. However, the means on this measure indicate participants *would* use both the real and virtual scenarios as a practice tool.

5.5.3 Analysis

The goals of a virtual interpersonal scenario can be met even if the virtual character is viewed as deficient on other measures. Results show that there are deficiencies in the virtual patient's authenticity, expressiveness and conversational behavior. Also, overall impressions of the virtual interaction were significantly lower than the real one. Yet the virtual interaction was viewed as equally educational, and students performed equally well on eliciting critical information from the patient.

Comments that summarized participant impressions include: "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 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." Educational goals were clearly met by the virtual interaction despite the system's deficiencies.

However, our goal is to eventually impact communication skills. The transition from a diagnosis training system to a communications skills training system will require substantial investment in improving conversation flow and expressiveness.

5.6 Post-Study Reflections

These results should be considered preliminary because of study attributes and several key differences between the groups.

Sample Size: The population size, particularly in the case of the SP experience (n=8), is too limited. A larger study was recently conducted, and results are currently being analyzed.

Participant Experience: For scheduling reasons, it was difficult to recruit medical students of equal experience levels. This difference likely affected study results. One effect could be that students with less experience do not yet know what a good or bad interview with an SP is. They may apply a different grading standard than the more experienced students. Another concern is that more experienced students likely conduct better interviews.

Different Institutions: Logistical issues made it difficult to run Group VP and SP at the same institution. It is possible students at each institution have a tendency (and/or training) to ask different questions. It is also possible that students at different institutions might rate standardized patients differently.

VP Voice Fidelity: As part of a separate study, some Group VP participants spoke to a patient with a computer-generated voice, and others spoke to a patient with pre-recorded responses from a real person. No significant difference was found between the text-to-speech and real-speech conditions on all measures [8]. Therefore, the two groups are combined together in our analysis. However, as VP expressiveness has been identified as affecting results, recorded professional talent might be necessary.

6 Conclusions and Future Work

Using the interaction between a medical student and a standardized patient, a study was conducted that compares the interaction between a virtual and real interpersonal scenario. Results show the virtual patient was not nearly as expressive as the standardized patient. This contributed to differences in the conversational flow and less rapport with the virtual patient. However, the virtual interaction was found to be similar to the real interaction on many important education measures. Participants elicited the same information from both virtual and standardized patients, and performed equally well overall. Furthermore, participants rated both interactions as equally valuable educational experiences. Finally, direct (big-picture) and indirect (subcomponent) measures of patient authenticity yielded conflicting results. Direct measures showed the real scenario was

more authentic, but indirect measures suggest – on a component level – the virtual scenario was similar to the real scenario.

These similarities and differences provide strong evidence that:

- The authenticity of virtual interpersonal scenarios should be assessed with a battery of indirect measures.
- The goals of a virtual interpersonal scenario can be met even if the interaction is not equivalent to the real one on other measures.
- Virtual characters in interpersonal scenarios must be expressive and significant effort should be put towards achieving this goal.

A larger study is planned to explore these issues more closely. New indirect measures of authenticity will be developed and tested for validity. The primary focus will be on making the virtual character more expressive. Although the high-level goals of the scenario were met, participants did not develop rapport with the patient, an important skill for a medical professional. A more expressive virtual character would improve the flow of conversation, elicit more natural behavior from the student, and improve rapport between the virtual patient and student.

ACKNOWLEDGEMENTS

We would like to thank Rebecca Pauley, Margaret Duerson and the rest of the staff of the Harrell Professional Development and Assessment Center at the University of Florida. We would also like to thank the Medical College of Georgia for the use of their facilities. A special thank you goes to our standardized patient Maria Martinez, as well as Emily Gorovsky for her help with editing this paper, narrating the video and providing the voice of DIANA. Lastly, we acknowledge the University of Florida Alumni Graduate Fellowships for partially funding this work.

REFERENCES

[1] N. E. Alessi and M. P. Huang, "Evolution of the Virtual Human: From Term to Potential Application in Psychiatry," *CyberPsychology & Behavior*, vol. 3, pp. 321-326, 2000.

[2] N. I. Badler, C. A. Erignac, and Y. Liu, "Virtual Humans for Validating Maintenance Procedures," *Communications of the ACM*, vol. 45, pp. 56 - 63, 2002.

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

[4] 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," in Proc. of The 7th Annual International Workshop on PRESENCE, Valencia, Spain, 2004.

[5] 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, pp. 583-598, 2001.

[6] M. Billger, "Colour Appearance in Virtual Reality: A Comparison between a Full-Scale Room and a Virtual Reality Simulation," in Proc. of The 9th Congress of the International Colour Association, 2001.

[7] J. L. Coulehan and M. R. Block, *The Medical Interview: Mastering Skills for Clinical Practice*, 3 ed. Philadelphia: F.A. Davis Company, 1997.

[8] R. Dickerson, K. Johnsen, A. Raij, B. Lok, T. Bernard, A. Stevens, and D. S. Lind, "Virtual Patients: Assessment of Synthesized Versus Recorded Speech," in Proc. of Medicine Meets Virtual Reality (MMVR) 14, Long Beach, CA, 2006.

[9] R. Dickerson, K. Johnsen, A. Raij, B. Lok, J. Hernandez, and A. Stevens, "Evaluating a Script-Based Approach for Simulating Patient-Doctor Interaction," in Proc. of The International Conference on Human-Computer Interface Advances for Modeling and Simulation, New Orleans, Louisiana, 2005.

[10] P. M. G. Emmelkamp, M. Krijn, A. M. Hulsbosch, S. de Vries, M. J. Schuemie, and C. A. P. G. 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.

[11] M. Garau, M. Slater, S. Bee, and M. A. Sasse, "The Impact of Eye Gaze on Communication Using Humanoid Avatars," in Proc. of The SIGCHI Conference on Human Factors in Computing Systems, Seattle, Washington, United States, 2001.

[12] 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 Journal)*, vol. 17, 2003.

[13] 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," in Proc. of IEEE Virtual Reality 2005 (VR 2005), Bonn, Germany, 2005.

[14] F. Lang, R. McCord, L. Harvill, and D. S. Anderson, "Communication Assessment Using the Common Ground Instrument: Psychometric Properties," *Family Medicine*, vol. 36, pp. 189-198, 2004.

[15] 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," *The Journal on Information Technology in Healthcare*, vol. 2, pp. 399-412, 2004.

[16] C. Nass and Y. Moon, "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues*, vol. 56, pp. 81-103, 2000.

[17] S. L. Oviatt, P. R. Cohen, M. Wang, and J. Gaston, "A Simulation-Based Research Strategy for Designing Complex Nl Systems.," in Proc. of ARPA Human Language Technology Workshop, Princeton, N.J., 1993.

[18] 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, pp. 68-78, 2002.

[19] 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," *Journal of Consulting and Clinical Psychology*, vol. 68, pp. 1020-26, 2000.

[20] 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, pp. 37-51, 2000.

[21] K. R. Thórisson, "Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People," in Proc. of The First ACM International Conference on Autonomous Agents, Marina del Rey, California, 1997.

[22] K. R. Thórisson and J. Cassell, "Why Put an Agent in a Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue," in Proc. of Lifelike Computer Characters '96, Snowbird, Utah, 1996.

[23] M. Usoh, E. Catena, S. Arman, and M. Slater, "Using Presence Questionnaires in Reality," *Presence: Teleoperators and Virtual Environments*, vol. 9, pp. 497-503, 2000.

[24] V. Vinayagamoorthy, A. Steed, and M. Slater, "Building Characters: Lessons Drawn from Virtual Environments," in Proc. of Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop, Stresa, Italy, 2005.

[25] 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, pp. 39-44, 2004.

[26] D. Wuillemin, G. van Doorn, B. Richardson, and M. Symmons, "Haptic and Visual Size Judgements in Virtual and Real Environments," in Proc. of The First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'05), 2005.