Kyle Johnsen*
Robert Dickerson
Andrew Raij
Cyrus Harrison
Benjamin Lok
Computer and Information Science
and Engineering
University of Florida
Gainesville FL 32607

Amy Stevens
D. Scott Lind
VA Hospitals
University of Florida

*Correspondence to
kjohnsen@ufl.edu

Evolving an Immersive Medical Communication Skills Trainer

Abstract

This paper presents our experiences in evolving the Virtual Objective Structured Clinical Exam (VOSCE) system. This system allows medical students to experience the interaction between a patient and a medical doctor using natural methods of interaction with a high level of immersion. These features enable the system to provide training on medical communication skills. We discuss the experiences of a group of medical and physician assistant students that pilot tested the system. Further, we examine the impact of evolving the system based on their feedback. The VOSCE system's performance in subsequent studies has indicated that end-user feedback improvements have significantly impacted overall performance and efficacy.

1 Overview

Doctor, I have a pain in my side! Please make it go away!

This common patient complaint is known as acute abdominal pain (AAP) and is a basic condition that medical students must be able to quickly and correctly diagnose. See Figure 1. The first step to diagnosis is not administering tests or interpreting test results, but rather the critical patient-doctor interview. F1

Educating these skills is a time-consuming process. First, students learn procedure and the appropriate questions to ask from textbooks. Second, students employ role-playing with instructors and fellow students to practice and hone these skills. Finally, they are evaluated using standardized patients, actors trained to simulate the symptoms of an illness. Standardized patients are currently the best option available to train, teach, and test interview skills. However, standardized patients can be limited in providing experience diversity, repeatability, quality control, and effectiveness. We believe a virtual patient would help address the first three issues and would be a powerful education tool to augment medical students' standardized patient coursework.

Researchers have begun exploring the benefits of simulating social situations. USC's CARTE group has applied intelligent agents to pedagogical and training applications (Rickel & Johnson, 1999), and the ICT group has created experiences to train military personnel in interpersonal leadership (Hill et al., 2003). Pertaub, Slater, and Barker (2001) observed participants with a fear of public speaking who spoke to an audience of virtual characters. Participants responded similarly when they spoke to an audience of real people. Further, experiencing a virtual social situation may reduce anxiety in reality. We aim to add to this research by investigating the following:

Presence, *Vol. 15, No. 1, February 2006*, 000-000

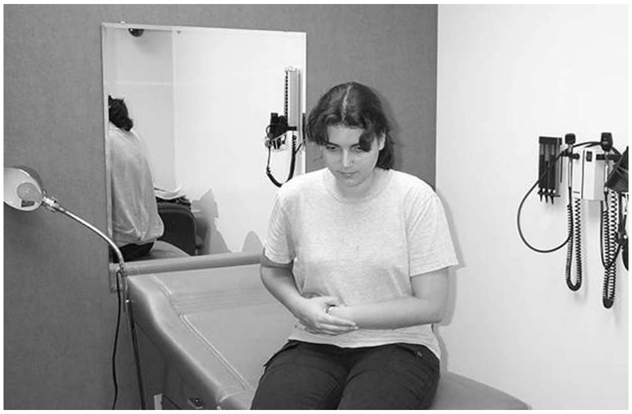

Figure 1. *A female standardized patient (trained actor) complains of abdominal pain.*

- For a highly constrained scenario, can VR systems train students in communication skills?
- What is the impact of natural interaction and a high level of immersion on the effectiveness of a VR system?
- How can a system overcome the technical limitations to provide a compelling interpersonal scenario?

To address the above questions, we aimed to create a compelling interpersonal experience using the best, and most cost-effective, technologies for immersion and natural interaction (e.g., display, speech recognition, gesture recognition, virtual character models). The resulting immersive Virtual Objective Structured Clinical Exam (VOSCE) system uses virtual characters that simulate the standardized patient experience. We installed the VOSCE system in a real examination room at the Harrell Adult Development and Testing Center at the University of Florida (UF). The Harrell Center is where UF medical students routinely interview standardized patients.

A data projector is used to display the virtual world (modeled as an extension of the real examination room) in the context of the real world. Head-tracking, gesture recognition, and speech recognition allow for natural interaction with the virtual world. The inhabitants of the virtual space are DIANA (Digital ANimated Avatar), playing the role of the standardized patient and VIC (Virtual Interactive Character) who represents an instructor.

We have run two studies with the VOSCE system. In a pilot study, medical students of varying experience interviewed DIANA and were evaluated by VIC. Modifications were then made to the system based on the feedback obtained from the pilot study. A second study was then run using only second year medical students. We present the results of those studies and the lessons learned.

2 Previous Work

2.1 Medical Training

Virtual characters have been applied as interactive agents for training and procedure planning. The Just VR system uses an immersive approach to have students experience and react to health emergency scenarios (Ponder et al., 2002). Similar in spirit to this work, Research Triangle Institute's (RTI) Virtual Standardized Patient is a commercial virtual character system for training medical students (Hubal, Kizakevich, Merino, & West, 2000). Students interact with 3D virtual patients displayed on a monitor using natural speech (with natural language processing), mouse, and keyboard.

2.2 Virtual Characters

Creating useful or believable agents has been a goal for artificial intelligence and simulation research for such applications as simulating military forces (Stytz & Banks, 2003) and crowds (Hamagami & Hirata, 2003). MIT's Synthetic Characters group is investigating the use of artificial intelligence to drive interactive autonomous agents (Burke, Isla, Downie, Ivanov, & Blumberg, 2001). Virtual characters have also been used as interactive guides (Thórisson, 1997). There are software libraries, such as Microsoft Agents, that add interactive personalities to application interfaces.

Highly detailed human models, complete with physical constraints, can be integrated with virtual parts, vehicles, and tools to evaluate ergonomics and usability. The Human Modeling and Simulation Group at the University of Pennsylvania use virtual humans for task analysis and assembly validation (Badler, Erignac, & Liu, 2002). Safework, EDS's Jack system, and Boston Dynamics are commercial packages that incorporate virtual humans for ergonomics and human factors analysis. Finally, VQ Interactive's BOTizen and Haptek Inc.'s HapPlayer software provide virtual characters for creating interactive and memorable websites.

2.3 Immersion

The link between interaction, immersion, and cognition is currently being researched. Insko's studies (2001) showed improved memory recall through natural locomotion and haptic feedback. Other studies suggest that natural locomotion improves higher order cognitive performance (Zanbaka et al., 2004), and naturally interacting with real objects improves problem solving task performance (Lok, Naik, Whitten, & Brooks, 2003).

3 Project Description

3.1 Project Goals

F2 Figure 2 shows the interactive virtual patient DIANA and the instructor VIC. These characters combine high-quality rendering, animation, and display, with speech and gesture recognition. The virtual characters provide scripted responses to the student's speech and gestures. With its high level of immersion and interactivity, this system allows participants to go through compelling experiences that help educate medical communication skills. The repetitive practice and potential availability of this technology has real merit in training and documenting physician competency before real patient interactions. Ultimately, the development of more varied and less common patient encounters will make this system a powerful educational tool. In the future this tool could be part of the standardized testing process.

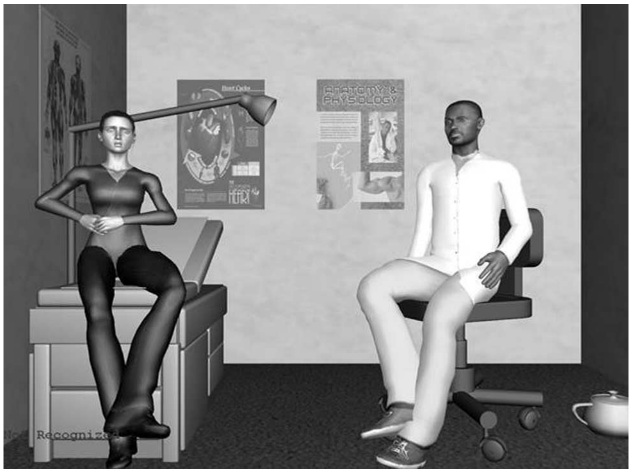

Figure 2. *DIANA, a female patient (left) complains of abdominal pain. VIC, the instructor (right) coordinates the diagnosis.*

3.2 Driving Application: Patient-Doctor Interaction

The interaction between patients and doctors during initial condition diagnosis is an interpersonal scenario 1) which can be simulated despite the limits of current technology, 2) where practice is costly, 3) where immersion and fidelity is important, and 4) which benefits from immediate feedback. Currently, high-level components of the patient-doctor interaction, such as bedside manner and grief counseling, would be severely hampered by even slight compromises in fidelity (inevitable with current technology). However, educating the students on the core interview process of

1. Determining key questions
2. Asking questions in the correct manner
3. Coming up with a list of possible diagnoses

imposes less of a requirement on the system to be perfect.

Patient-doctor interaction requires substantial first-hand experience to become proficient. Because of logistical difficulties, there are insufficient opportunities to expose students to an adequate variety of scenarios. This results in many medical students not receiving sufficient practice in communication skills before they reach a real patient's bedside.

Communication skills are an important part of the medical school curriculum. Research has shown communication skills can be taught and practiced and do not just improve over time with clinical experience (Lang, McCord, Harvill, & Anderson, 2004). To provide maximum benefit, the experience should simulate, as closely as possible, the actual scenario.

An informal survey of a sample of third year Shands Hospitals medical students highlighted that although students feel they get adequate instruction in communication skills, they receive little feedback—none of it immediate—on their performances. The design of the virtual scenario includes a virtual instructor who can address these issues of feedback.

3.3 Acute Abdominal Pain (AAP) Diagnosis

Acute abdominal pain (AAP) is defined as follows:

> The term *the acute abdomen* refers to the presence of an acute attack of abdominal pain that may occur suddenly or gradually over a period of several hours. The patient with this symptom complex may confront the surgeon, internist, pediatrician, and obstetrician, creating a problem in clinical diagnosis requiring an immediate AQ: 1 or urgent decision regarding the etiology [cause] and method of treatment. (Diethelm, 1997)

AAP is one of the most common ailments encountered by doctors and is a basic scenario in patient-doctor interaction and communication skills education. The doctor begins diagnosis by asking the patient a series of questions about the pain (history of present illness or HPI). Sample questions include “What brought you into the clinic today?” “How long have you had the pain?” and “On a scale from 1 to 10, please rate the pain.” In AAP diagnosis, doctors should ask the proper questions and determine the correct diagnosis. The traversal of questions is a very well-defined process.

The students are graded on how many of the eleven key questions they asked. These questions must be asked to ascertain the correct diagnosis. The common mistakes are omitting questions, incorrectly asking questions, or not evaluating the nonverbal cues from the patient’s gestures and verbal responses (e.g., recurring cough).

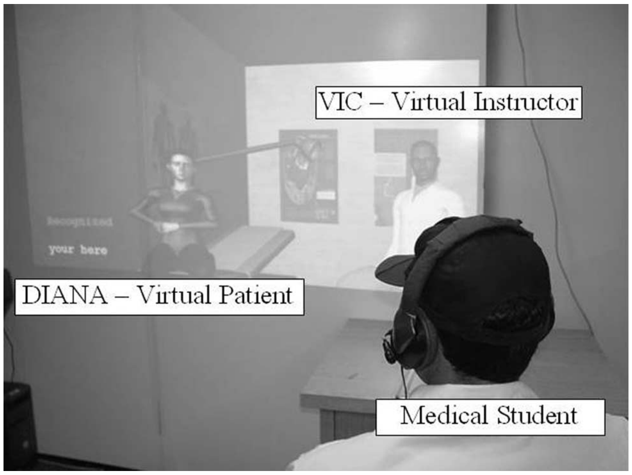

Figure 3. *The (right) student is diagnosing (left) DIANA, a patient with acute abdominal pain, while (middle) VIC observes. The colored headset is for head tracking.*

3.4 Scenario

The VOSCE system models a typical standardized patient AAP scenario. A 19 year- old Caucasian female with a complaint of acute abdominal pain is in an examination room with a medical student. The student needs to ask the patient relevant questions to come to a diagnosis of appendicitis. DIANA plays the role of the patient with appendicitis. VIC is also in the room playing the role of an observing expert (Figure 3). DIANA and F3 VIC’s scripts, which included gestures and audio responses, were created in consultation with several teaching medical faculty and students with substantial standardized patient experience.

4 VOSCE System Version 1

4.1 Overview

The VOSCE system is composed of the following (Figure 4): F4

- 2 PCs (networked)

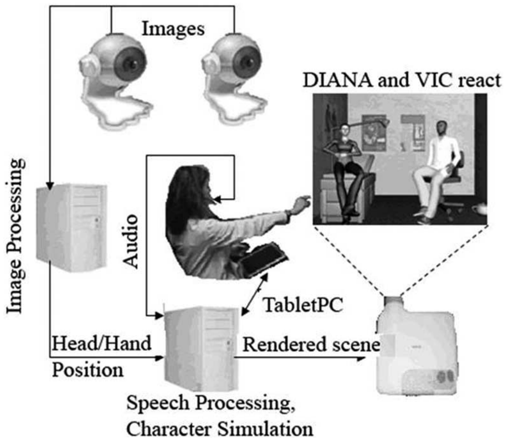

Figure 4. *System layout.*

- Data projector to display the virtual characters at life-size
- 2 web cameras to track the user's head and hand
- TabletPC
- Microphone

The prototype system cost less than $7,000. In our system, the medical student asks a single question and/or makes a gesture, which the virtual character perceives, thinks about, and then visually and audibly responds to.

4.2 Virtual Character Perception

The student uses speech and gestures to interact with DIANA and VIC. The system receives audio and video input from the microphone and web cameras. The audio is processed into phrases by a commercial speech recognition engine (Dragon Naturally Speaking 7 Professional). The system displays the most recent recognized parse on the phrase. AQ: 2

The HPI (history of present illness) portion of the exam consists of a set of questions which the students are taught to ask. The query database contains the most likely forms of each question, and several questions could map to the same response. For example, "Are you nauseous?" and "Have you been vomiting?" both cause DIANA to tell the student that she has felt sick to her stomach. The system uses a heuristic (described in Dickerson et al., 2005) to match the recognized speech to a particular question.

The system tracks the 3D trajectory of the students' hand with a marker-based tracking algorithm (Jackson et al., 2004). Two gestures were recognized, handshaking and pointing. Handshaking is signaled if the student held their hand in front of their body for more than two seconds. Pointing is detected by finding the intersection of a ray (from the tracked head to the hand) and objects in the scene. A "laser pointer" red dot appeared where the system determined the student was pointing.

Tracking the student's head position enables DIANA and VIC's eyes to focus on the student. Correct perspective warping (Raskar, 2000) of the rendered image emphasizes the characters' gaze directions, and maintains the illusion of the virtual examination room as an extension of the real room.

4.3 Virtual Character Cognition

The system used a simple state-based approach that transitions between actions depending on input (speech and gestures) from the perception stage. Transition rules are based on accepted medical doctrine for the scenario. Actions include the virtual character speaking statements, changes in emotion, or animation. Our medical collaborators created DIANA and VIC's responses and verified that acute abdominal pain diagnosis training lent itself well to this architecture.

4.4 Virtual Character Response

DIANA and VIC are displayed at life-size using data projectors. This research proposes that seeing a human face and form at the appropriate size (as opposed to on a monitor) increases immersion and triggers psychological responses. The system uses Haptek Inc.'s character animation library, which can generate high-quality, dynamic facial expressions and gestures. The characters verbal responses were driven by the doctor-designed scenario and audio generated by AT&T Wizard's Natural Text-To-Speech software.

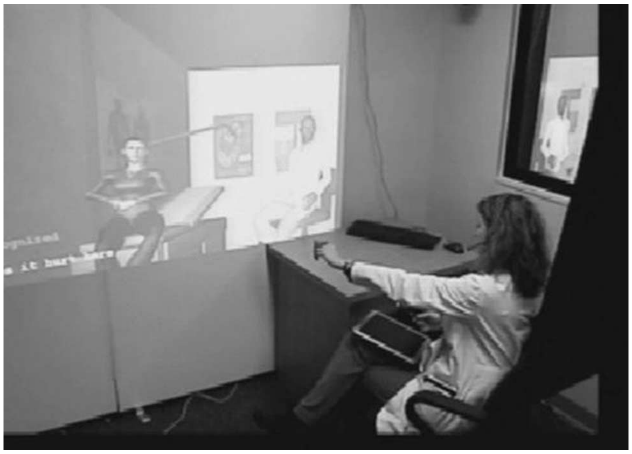

Figure 5. *Student asks, "Does it hurt here?"*

4.5 Immersion

VOSCE was installed at the Harrell Adult Development and Assessment Testing Center, where students routinely go to meet with standardized patients. The exam rooms have closed circuit recording for real-time observation and to record each session. Standardized patients await students in the center's examination rooms. DIANA and VIC were in Examination Room #3. We feel that this contributes to immersion by providing a *familiar,* realistic environment as opposed to a computer science laboratory.

We were interested in how much of an education value-add would be realized if the system provides natural interaction and a high level of immersion. For example, one question in AAP diagnosis is, "Does it hurt here?" where the doctor points either to a place on herself or
F5 on the patient (Figure 5). Does it help to more closely simulate the real experience, *i.e.,* instead of moving a cursor and clicking to select a location, that the student actually points at DIANA?

5 Study I

5.1 Design

In an AAP standardized patient scenario, the medical student is given basic information about the patient before entering the room. They then enter the room, greet the patient, and obtain the history of present illness (HPI). To determine the HPI, the student needs to ask a prescribed set of questions to find out why the patient has come into the clinic. Asking a question can consist of *both* a verbal and gesture component. A physical exam usually follows. This study focused on *only* the greeting and HPI component.

Medical students conducted AAP diagnosis with DIANA and VIC to evaluate if we had achieved our goal of developing a training, teaching, and evaluation tool. Further, we looked to evaluate the impact of the various immersive system components, such as gestures and speech, as well as the much requested feedback and instruction provided by VIC.

5.2 Procedure

5.2.1 Pre-Experience. After arriving at the Harrell Center, each participant signed a consent form. The participant then created a voice profile which took ten minutes. We do not consider this to be a major inconvenience because each student need do this only once, and the saved profile could be used for a variety of scenarios. The participant is then led to another room to fill out a background survey on prior exposure to standardized patients, abdominal pain scenarios, and medical examinations. Meanwhile, the virtual character system was started. The participant was then brought to the examination room where DIANA's patient information chart was displayed on a TabletPC in the door basket. They picked up the TabletPC, reviewed the information, and then entered the room to meet DIANA and VIC.

5.2.2 Experience. VIC began the experience by explaining some information regarding the scenario. He also reminded the participant how to perform the gestures, how to interrupt DIANA if she did not understand correctly, and how to end the session early. VIC then asked them to begin their examination of DIANA. After eight minutes, VIC interrupted the session and announced that two minutes remained. After ten minutes, VIC ended the session and asked the participant for their differential diagnosis.

Finally, VIC reported which, if any, of the eleven key questions were not asked by the participant. VIC then explained the correct differential diagnoses. This completed the experience, and the student left the examination room. From start to finish, DIANA and VIC were always present.

5.2.3 Post-Experience. After the examination, the participant filled out a presence (Usoh, Catena, Arman, & Slater, 2000) and copresence (Mortensen et al., 2002) questionnaire. Two other measures were also used. One was a survey used to measure the performance of real standardized patients (Wind, Van Dalen, Muijtjens, & Rethans, 2004) which was adapted to evaluate DIANA. The other gauged the importance of the system components and evaluated the system as a whole. Finally, in an oral debriefing, the participant provided qualitative feedback.

5.3 Results

5.3.1 Population. A total of seven participants were involved in the study. There were three male and three female medical students (three third years and three fourth years) and one third year female physician assistant student. All had substantial prior experience with standardized patients (at least 5, average: 10–20). Five had experiences with standardized patients with AAP, and six had experiences with real patients with AAP. This indicated that the participants had significant prior experience in similar scenarios.

5.3.2 Task Performance. *Scoring:* Each participant was evaluated on the following criteria: 1) Correct greeting etiquette (Introduce self, shake hands, query for chief complaint), 2) Eleven core AAP diagnosis questions that need to be asked to obtain the correct diagnosis, and 3) Differential diagnosis (what is the final evaluation). Five out of seven students introduced themselves and shook hands. All queried for the chief complaint. Out of the eleven core questions, the average number asked was 6.4 (7.0 is a passing grade). Four out of seven received passing marks. Our medical collaborators verified that this was a typical result.

All students forgot to ask DIANA to "tell them more about the pain," a common mistake, but important because patients often provide only simple answers until asked to elaborate.

After the HPI, there were six possible correct differential diagnoses. The most critical were appendicitis and ectopic pregnancy because these would require immediate surgery. One student included ectopic pregnancy, while four included appendicitis. Four students had one correct diagnosis and three had two. Only one included an incorrect diagnosis. We did not specifically ask them for all possible diagnoses so we consider this acceptable.

5.3.3 Technology Survey. This self-created survey was used to informally gauge the impact of technical components on the experience by asking questions about system features, not specifically the technology, to avoid possible confusion. Responses are on a 7-point Likert scale (1 = strongly disagree, 4 = neutral, 7 = strongly agree).

System components that increased the level of immersion were usually rated with highly positive responses. Participants indicated that interacting with DIANA and VIC with speech (mean: 6.7), seeing DIANA and VIC at life-size (6.3), and having the system at the Harrell Center (5.9) were very important to the experience. Also training (6.3) and testing (6.3) were seen as potential uses for the system. To use the system for evaluation (4.9), however, had mixed results.

We had some unexpected negative results. Interacting with DIANA using handshaking and pointing gestures was not viewed as important to the experience (3.0). The scene "moving when they moved their head" (perspective-correct rendering) was generally not thought critical to the experience (3.3), although debriefing comments were mixed. Interestingly, students responded to the possibility of typing their questions if it improved accuracy (3.9) either high or low, depending on their frustration with the accuracy of DIANA's responses.

5.3.4 Standardized Patient Satisfaction Survey. This survey is adapted from the survey used to evaluate standardized patients. Answers are on a 5-point Likert scale (1 = strongly disagree, 3 = neutral, 5 = strongly agree).

Students indicated DIANA appeared authentic (mean: 4.1), communicated how she felt during the session (4.3) and stimulated them to ask questions (4.1). Students reported mixed feelings about whether DIANA listened to them or not (3.0). The only clearly negative response was that students felt that DIANA did not answer questions in a natural manner (2.4).

Scores reported that VIC interrupting them was helpful (4.3) and his criticism was constructive (4.3). However, they liked only getting feedback at the end of the session (4.1).

Overall, on a scale from one to ten, DIANA was given an overall score of 6.4 for the interaction (authenticity, accuracy, and symptom display). For comparison, the average score for standardized patients is 7.47.

5.4 Discussion

5.4.1 Technology Survey. Gestures and perspective-correct rendering were viewed negatively by the students. We believe that the reason for this is that the scenario did not require them to be used. Also, noise in the optical tracking of the participant's head caused noticeable jitter in the display which students complained about in the debriefing.

Participants agreed life-size characters were very important. We believe this suggests that using data projectors to show full-body life-sized virtual characters (as opposed to the limited display space of monitors) was critical to the experience.

All students felt that the system would be an invaluable tool in training and testing, especially to those without much patient experience (students in their first two years). Feelings were more neutral on the system as a skills evaluation tool. This is understandable, as system errors and fidelity compromises are much more critical if used to evaluate performance.

5.4.2 Standardized Patient Satisfaction Survey. High scores for DIANA's authenticity and communication skills validate our virtual character approach, as well as the content of the scenario.

Work needs to be done to improve how DIANA responds. The lack of higher level information in text to speech, such as tone, may have made her responses seem simulated. VIC was highly praised, which correlated with the students' desire for feedback on their performance. This suggests finding more ways to incorporate VIC, such as context sensitive help and feedback on interruptions or nonverbal communication.

5.4.3 Debriefing. The debriefings yielded important comments, constructive criticisms, and positive feedback on various system components.

5.4.3.1 Instructor. VIC's presence enhanced the experience by providing helpful information during the session and post-session feedback.

> I liked that [VIC] gave me feedback at the end and told me exactly which questions I forgot to ask.

Several participants suggested VIC could provide hints when asked, particularly for first and second year medical students who are not experienced enough to get through an AAP scenario.

5.4.3.2 Speech. Speaking with DIANA and VIC enhanced the experience.

> I don't think of it as much as watching a computer screen as actually interacting with a person. You're actually talking to a patient; you're not typing in something and waiting for a response.

5.4.3.3 Speech Comprehension. DIANA answering questions incorrectly or repeating previous answers detracted from the experience. This made it harder to get diagnosis-critical information. For example, one person asked DIANA several times whether she had a history of gall bladder problems, but the query database for DIANA had no information about her gall bladder and thus the system responded incorrectly. This visibly frustrated the student. Some participants reported learning how to ask DIANA questions to avoid improper responses. Some even thought the flaw was a feature.

> I think it's good for us, too, because if the patient doesn't understand the question, which is inevitably going to happen in real life, too, it forces you to think about other ways to ask questions.

Others noted that it was quite distracting to them. Students would benefit from an improved speech recognition system.

> [I got] caught up with trying to think of a way to phrase [the question] rather than taking her history.

5.4.3.4 Scenario Content. From the debriefing we learned that DIANA had a tendency to offer too much information, which is not typical of most patients.

> Often times I actually thought she gave more detailed answers than real people. Another participant joked that he did not need to interview her after her initial complaint. Others made the point that some patients give up information more readily than others. One participant suggested simulating this variability by providing varying difficulty levels. DIANA, in her current state, might be considered an easy patient because she offers so much information. Harder difficulty level patients would provide less information.

5.4.3.5 Gestures. Most felt the gestures were not very useful, and many did not even remember to use them.

> I think the whole shaking hand thing and pointing is not really that important. Some said DIANA pointing to the right place on her abdomen indicated where her pain was, so they did not see a need for pointing at her. Some saw handshaking as a novelty while others saw no value in it because DIANA is not real.
>
> [People] would not accept an image as someone they can shake their hands with.

6 VOSCE System Version II

6.1 Overview

Given the feedback and survey results we received from the medical students, we set out to improve the system.

6.2 Script Improvement

Based on analysis of the interviews performed by the Study I students in Dickerson et al. (2005), we expanded the scope (what types of questions asked) and depth (variations on each question) a great deal. We showed that over 60% of the speech input was mapped to the correct response. Of the failed responses 51% of them were just things that weren't in the scope of the script and 21% were variations on a question that were not in the depth of the script. The rest were errors that are difficult to handle, such as summarization (when the doctor tries to repeat what the patient said to them) and empathetic statements. By incorporating the new questions asked by students in Study I as well as the variations on existing questions, we believe that the system will be able to handle a higher percentage of queries correctly.

6.3 Speech Understanding

The heuristic described in Dickerson et al. (2005) underwent minor adjustments. DIANA was made to respond less often to vague input (errors in speech recognition typically). While this means that DIANA responds less, it also means that when she does respond, she will respond correctly more often. We believe that this will improve the satisfaction of using the speech input system.

6.4 Tracking

The passive, colored marker-based tracking system used in the study jittered significantly or lost tracking altogether. The markers were also encumbering because of the relatively large size required by the use of the inexpensive web cameras. We believe that this was directly correlated to the students' acceptance of head-tracked rendering and gesture recognition.

Most digital cameras pick up infrared (IR) light fairly well. We took advantage of this and moved to an active marker-based system using IR LEDs and put infrared filters on the cameras. As a result the system now is lighting independent (important because the user is often in the path of the projected image), uses a much simpler and faster algorithm, and has far less jitter.

6.5 Scenario Changes

In an upcoming series of studies, we aim to compare medical students' performance with virtual patients and standardized patients. To this end, we took out the feedback that VIC gives students at the end of the experience. While students liked this, it is not something that is present in a standardized patient encounter. We choose to remove this potential confounding factor for a formal comparison between virtual and standardized patients. DIANA's responses were also shortened to make her a more realistic patient.

7 Study II

In December 2004, we conducted a follow-up study to identify the impact of the above changes.

7.1 Design

We randomly separated the study participants into two groups, one where the virtual characters used real speech (recorded human speech) and another where they used synthetic speech as in the pilot study to identify any differences between these conditions.

7.2 Results

We compared the results of Study II with that of Study I to see if there were any trends. While the studies were too different to say definitively if the differences are significant, there were some important general trends.

7.2.1 Population. The study population was restricted to second year medical students. Ten students—all with nearly identical backgrounds—participated in the study. There were 6 men and 4 women. They had all seen between 10 and 30 standardized patients and had done the AAP scenario with standardized patients between 4 and 6 times. We feel that this is the target group for VOSCE.

7.2.2 Task Performance. Half of the students introduced themselves. Eight out of ten students queried for the chief complaint. Out of the 11 core questions, the average number asked was 5 (7.0 is a passing grade). Only one out of ten received a passing grade. We believe the significant difference on number of relevant questions asked ($p = .03$) between this group and the participants from the first study is a result of the differences in experience between the two groups.

The lower experience levels of the students in the second study also showed itself through much lower frequency in asking questions about sensitive topics, like sexual history. Study II participants forgot more often to ask if the patient was sexually active ($p = .05$) and to ask when DIANA's last menstrual period was ($p < .01$). Surprisingly, all second year students gave a correct diagnosis of appendicitis. However, the lack of other possible diagnoses, such as ectopic pregnancy, is another consequence of the experience differences between the two study groups.

7.2.3 Technology Survey. Again, we had positive responses towards some of the immersive technology speech (mean: 6.6), life-size characters (5.8), and the Harrell Center (4.8), although none significantly different from the previous study. Using the pointing gesture this time around was seen as more important than last time (4.2 vs. 3.0, $p = .05$). There was also a trend towards head tracked rendering being more important (4.4 vs. 3.2, $p = .10$). Students felt that the accuracy of the speech recognition was more sufficient this study (4.9 vs. 3.9, $p = .10$).

Students again strongly agreed that the system is useful for training and teaching, and were again neutral on its use for evaluation. The scores were noticeably lower for each category than the last time, (5.9 vs. 6.3, 5.6 vs. 6.3, 4.2 vs. 4.9) but not statistically significant.

Five of the students indicated that they would use the system weekly and four said they would use it monthly. One student did not answer the question.

Students felt that DIANA in general spoke more naturally than in the previous study (5.5 vs. 4.6, $p = .09$). In the between subjects study between synthetic speech and real speech the difference was significant (6.2 vs. 4.8, $p = .05$).

7.2.4 Standardized Patient Satisfaction Survey. The most notable change from Study I was that students felt DIANA now answered questions in a more natural manner (3.7 vs. 2.4, $p = .02$). Surprisingly, the difference between the real speech and synthetic speech groups was not significant.

Participants also tended to rate DIANA higher on average (7.2 vs. 6.4 on a scale of 1 to 10, $p = .10$).

7.3 Discussion

7.3.1 Technology Improvements. Using active, IR markers made head tracking and the pointing gesture noticeably more stable, and thus easier to use. We believe this was reflected in the improvements in the technology survey scores for the second study. Both system factors, which are directly tied to the tracking system, scored higher. The improvement in the gesture component could also be attributed to removing the handshake gesture based on user feedback from the first study. We believe that further improvements to the tracking system are still necessary to remove jitter and allow for more unconstrained movement. We are exploring the use of inexpensive (<\$300) digital video cameras with a night vision mode which enhances their ability to pick up IR light.

The speech recognition component (Dragon Naturally Speaking) did not change for the study, yet students tended to think that the accuracy of the speech recognition was sufficient to complete the task as opposed to in Study I where they had mixed feelings. Adding more scope and depth to the script as well as improving the speech understanding component can explain the difference. While the recognition of speech was no different, the patient responded correctly more often. In this study 70% of the queries were responded to correctly as opposed to 60% in the last study.

7.3.2 Virtual Patient. Student opinion was that DIANA now responded more naturally to questions. We believe this to be related to three system changes: script improvements, speech understanding improvements, and using real speech for some of the interactions instead of synthetic speech. A *natural* response is not just the quality of the voice, but also what was said, which could explain why there was not a significant difference when only real speech vs. synthetic speech was examined.

Participants also tended to rate DIANA on a scale from 1–10 higher (7.2), much closer to the 7.47 that is the average for standardized patients. While direct comparisons between real and virtual experiences are questionable, anecdotally, students seem to treat DIANA as though she is a real patient. This was confirmed through the comments of numerous medical faculty who have watched recorded sessions of medical students with DIANA.

7.3.3 Debriefing *7.3.3.1 Educational Objectives.* All students expressed that there is educational value in interviewing DIANA. Practicing the process of forming a diagnosis was the most valuable aspect of the system.

> Sometimes just having it come out of your mouth is useful. This is a way to do that without having to have somebody there.

Without any knowledge of the previous study where VIC provided feedback, at least one student specifically mentioned that the system would be better if feedback was provided.

> I felt like the diagnosis should be given at the very end. Right now we dont get that feedback with standardized patients.

This further confirms that feedback should be one of the critical components of the VOSCE system.

7.3.3.2 Gestures. Students still had many criticisms of the pointing gestures, but all tried to use them. Some felt that it was hard to point in the right direction, and that they were too far away from DIANA to use it appropriately. Many revealed that they thought pointing was completely unnecessary for the experience.

7.3.3.3 Speech Recognition. Students were impressed by DIANA's ability to answer most of the questions spoken to her.

> Understanding most of what I said was really neat.

Students this time around complained about the system not being able to support compound sentences and having to speak very clearly to the system.

> I often had to repeat what I said and try to enunciate a little clearer. The more complex my sentences were the less likely they were to get it also. I had to make really short non-compound sentences.
>
> I felt like I had to speak more clearly.

Our medical collaborators noted the older students (in the first study) did not mention this because they know that asking compound sentences is bad in practice and that speaking clearly is critical.

Currently, DIANA says nothing when she doesn't understand a question (she did this much more often than the first study to limit the amount of information she revealed by mistake). Most students felt this silence was unnatural and suggested she respond in some way to at least acknowledge that something was heard.

7.3.3.4 Scenario Content. While some students felt that DIANA's posture and disposition were appropriate for the scenario, the overwhelming opinion was that it needs to change. This was not seen as negatively as in the first scenario.

> She should have doubled over, squirmed, hunched over, laying down . . .
>
> She didn't really act like she was in pain. She was talking with a regular voice . . . usually patients are hunched over; it's hard for them to speak.
>
> With standardized patients you know that they're not feeling too good because they're lying, and holding their stomach.

AQ: 3

A different posture, improved voice acting and animations should improve DIANA's realism.

7.3.3.5 Synthetic Speech vs. Real Speech. Students who tried DIANA with synthetic speech felt DIANA was not expressive enough.

> The biggest thing was her voice was different. I knew it was artificial . . . I couldn't tell any feelings in her voice. No affect in her voice, which might be something important when dealing with a patient.

8 Conclusion and Future Work

We believe that in creating an immersive virtual character system with natural interaction, we have achieved our goal of more fully exploiting the capabilities of virtual characters. The most important part of our system is the fact that it is being continually refined through the use of real user feedback. By making focused improvements to the content and technology after an initial pilot study, a positive effect was seen on user satisfaction in a subsequent study. The performance of our virtual patient, DIANA, was given a high mark in the pilot study (6.4 out of 10 on the standardized patient evaluation questionnaire), and an even higher mark, 7.2, in the follow-up study. Participants maintained that life-sized virtual characters, speech recognition, and having the system at the Harrell Center were crucial while gesture support is improving but needs work to be accepted.

Given the overall positive feedback on the system, many further studies are planned. We are looking into exploring the effect of real speech vs. synthetic speech more closely. A large controlled study between a standardized patient and a virtual patient trained with the same script is also planned. Further studies will look into the effect of incorporating more nonverbal communication into DIANA's responses and actions. In addition, the system will continue to improve in content and technology based on user feedback.

We hope the lessons learned and experiences gained will provide insight to developers of similar projects. Understanding how to interact with virtual characters is the critical first step to better realize their potential for educating interpersonal skills. Obtaining user feedback is critical to that understanding.

References

Badler, N., Erignac, C, & Liu, Y. (2002). Virtual humans for validating maintenance procedures. *Communications of the ACM, 45*(7), 56–63.

Burke, R., Isla, D., Downie, M., Ivanov, Y., & Blumberg, B. (2001). CreatureSmarts: The art and architecture of a virtual brain. *Proceedings of the Game Developers Conference,* San Jose, CA, 147–166.

Dickerson, R., Johnsen, K., Raij, A., Lok, B., Hernandez, J., & Stevens A. (2005). Evaluating a script-based approach for simulating patient-doctor interaction. To appear in *2005 International Conference on Human-Computer Interface Advances for Modeling and Simulation.* AQ: 4

Diethelm, A. (1997). The acute abdomen. In D. Sabistan (Ed.), Textbook of surgery: The biological basis of modern surgical practice (chap. 29). 15th Edition. Philadelphia: Saunders.

Hamagami, T., & Hirata, H. (2003). Method of crowd simulation by using multiagent on cellular automata. *Intelligent Agent Technology. 2003 IEEE/WIC International Conference,* 13–16 Oct. 2003, 46–52.

Hill, R. Gratch, J., Marsella, S., Rickel, J., Swartout, W., & Traum, D. (2003). Virtual humans in the mission rehearsal exercise system. *Kynstlich Intelligenz, 17*(4), 32–38.

Hubal, R., Kizakevich, P., Merino, D., & West, S. (2000). The virtual standardized patient: Simulated patient-practitioner dialogue for patient interview training. In *Envisioning Healing. Interactive Technology and the Patient-Practitioner Dialogue,* J. D. Westwood, H. M. Hoffman, G. T. Mogel, R. A. Robb, & D. Stredney (Eds.). Amsterdam: IOS Press.

Insko, B. (2001). *Passive haptics significantly enhances virtual environments.* Unpublished PhD dissertation, Department of Computer Science, UNC-Chapel Hill, Chapel Hill, North Carolina.

Jackson, J., Lok, B., Kim, J., Xiao, D., Hodges, L., & Shin, M. (2004). *Straps: A simple method for augmenting primary tracking systems in immersive virtual environments.* (Future Computing Lab Technical Report.) University of North Carolina at Charlotte.

Lang, F., McCord, R., Harvill, L., & Anderson, D. (2004). Communication assessment using the common ground instrument: Psychometric properties. *Family Medicine, 36*(3) 189–198.

Lok, B., Naik, S., Whitton, M., & Brooks, F. (2003). Effects of interaction modality and avatar fidelity on task performance and sense of presence in virtual environments. *Presence: Teleoperators and Virtual Environments,12*(6), 615–628.

Mortensen, J., Vinayagamoorthy, V., Slater, M., Steed, A., Lok, B., & Whitton, M. (2002). Collaboration in tele-immersive environments. In *Proceedings of the Eighth Eurographics Workshop on Virtual Environments,* May 30–31.

Pertaub, D., Slater, M., & Barker, C. (2001). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments, 11*(1), 68–78.

Ponder, M., Herbelin, B., Molet, T., Scherteneib, S., Ulicny, B., Papagiannakis, G., et al. (2002). Interactive scenario immersion: Health emergency decision training in JUST project. *VRMHR2002 Conference Proceedings.*

Rainer, S., & Jie, Z. (2002). Head orientation and gaze direction in meetings. *CHI '02 extended abstracts on human factors in computing systems.* April 20–25, 2002, Minneapolis, Minnesota.

Raskar, R. (2000). Immersive planar displays using roughly aligned projectors. *Proceedings of IEEE Virtual Reality 2000.*

Rickel, J., & Johnson, W. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence, 13,* 343–382.

Stytz, M., & Banks, S. (2003). An architecture to address uncertain requirements and composability for intelligent agents in distributed simulations. *Design and Application of Hybrid Intelligent Systems* (pp. 749–758). Amsterdam IOS Press.

Thórisson, K. (1997). Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people. *Proceedings of the ACM First International Conference on Autonomous Agents,* 536–537.

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments, 9* (5), 497–503.

Wind, L., Van Dalen, J., Muijtjens, A., & Rethans, J. (2004). Assessing simulated patients in an educational setting: The MaSP (Maastricht assessment of simulated patients). *Medical Education, 38,* 39–44.

Zanbaka, C., Lok, B., Babu, S., Xiao, D., Ulinksi, A., & Hodges, L. (2004). Effects of travel technique on cognition in virtual environments. *Proceedings of IEEE Virtual Reality 2004,* Chicago, 149–156, 286.