The Validity of a Virtual Human Experience for Interpersonal Skills Education

Kyle Johnsen†, Andrew Raij†, Amy Stevens, MD‡, D. Scott Lind, MD*, Benjamin Lok, Ph.D†

†Computer Information Science and Engineering
University of Florida
Gainesville, FL 32611
{kjohnsen,raij,lok}@cise.ufl.edu

‡College Of Medicine
University of Florida
Gainesville, FL 32611
Amy.Stevens@va.gov

*Department of Surgery
Medical College of Georgia
Augusta, GA 30912
dlind@mail.mcg.edu

ABSTRACT

Any new tool introduced for education needs to be validated. We developed a virtual human experience called the Virtual Objective Structured Clinical Examination (VOSCE). In the VOSCE, a medical student examines a life-size virtual human who is presenting symptoms of an illness. The student is then graded on interview skills. As part of a medical school class requirement, thirty three second year medical students participated in a user study designed to determine the validity of the VOSCE for testing interview skills. In the study, participant performance in the VOSCE is compared to participant performance in the OSCE, an interview with a trained actor. There was a significant correlation ($r(33)=.49$, $p<.005$) between overall score in the VOSCE and overall score in the OSCE. This means that the interaction skills used with a virtual human translate to the interaction skills used with a real human. Comparing the experience of virtual human interaction to real human interaction is the critical validation step towards using virtual humans for interpersonal skills education.

Author Keywords

virtual humans, virtual characters, virtual reality, validation, medicine, multimodal interfaces.

ACM Classification Keywords

H.1.2 Models And Principles: User/Machine Systems, H.5.2 Information Interfaces And Presentation: User Interfaces, I.6 Simulation And Modeling: Applications, J.3 Life and Medical Sciences

INTRODUCTION

Virtual human experiences may one day be ubiquitous in education. When using real humans is difficult, impossible, or dangerous, virtual humans may serve as substitutes. Before this can happen, it must be validated that when using virtual humans, the important educational objectives are met. An important objective in medical education is evaluating clinical examination interview skills. Experts evaluate medical students while the students perform Objective Structured Clinical Examinations (OSCEs)[21]. In an OSCE, a medical student conducts an interview with a hired actor called a standardized patient (SP). The SP simulates a real patient. We created and refined a virtual human experience in which a student can perform a Virtual OSCE (VOSCE)[14, 15]. In the VOSCE, a virtual human (VH) simulates a real patient. The results of a formal user study demonstrate that student performance when interviewing a VH in the VOSCE correlates to performance when interviewing an SP in the OSCE. This paper validates the VOSCE for testing clinical examination interview skills.

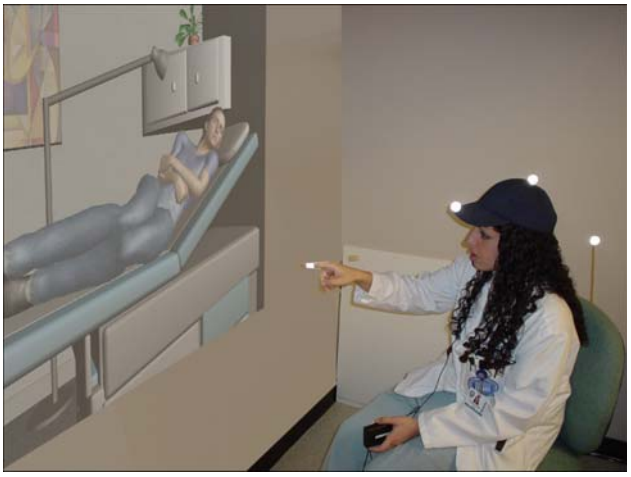

Figure 1. A student interacts with the VH during the VOSCE. Retroreflective tracking markers placed at various locations to track head gaze, pointing, and body lean. The flash from the camera illuminates the markers.

In order for performance in the VOSCE and OSCE to be compared, students must be able to interact with the VH as they would with an SP. The interaction skills set needed to interact successfully must be similar. As a result, a natural and transparent interface with the VH is required. A student is shown during a VOSCE in Figure 1. The VH is projected on an examination room wall at life-size. The student interacts with the VH as they would with a real human, using gestures and speech. The VH interacts with the student in the same way.

The metric used to evaluate a medical student's interview performance is a checklist of required interview skills. A student using good interview skills obtains all of the information required for an accurate diagnosis, does so efficiently, in the correct order, and follows proper patient-doctor etiquette. In pilot studies with the VOSCE, experts noted *that it was evident which students possessed adequate interview skills and which did not*. In the user study described in this paper, an expert uses the interview skills checklist to evaluate the interview skills of 2nd year medical students in both the VOSCE and OSCE.

We present data showing that student performance on the VOSCE, as evaluated by a medical expert, is significantly correlated with student performance on the OSCE. *The study design and the correlation in student performance validate that the VOSCE can be used to evaluate medical students' interview skills in clinical examinations*. Comparing the experience of virtual human interaction to real human interaction is the critical validation step towards using virtual humans for interpersonal skills education.

PREVIOUS WORK

Virtual Human Technology

Researchers have worked to establish the requirements of realistic virtual humans. Badler et al [2, 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." Thórisson and Cassell [25] 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 [26] 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 and Moon [18] have 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.

Our work is different from this work, in that we attempt to show where virtual humans can currently be successful in real world applications. We research the important characteristics for *effective* virtual humans, not *realistic* virtual humans.

Virtual Human Applications

The four primary application fields of virtual humans are the military, medicine, psychology, and entertainment. However, very little of this work has been validated.

In military simulations virtual humans are combatants, civilians, and fellow team members. USC's ICT group has applied interactive virtual human technology to military leadership training in its Mission Rehearsal Exercise trainer[9]. In this system, users interact naturally with virtual humans projected on a larger than life display.

Research is also being conducted in the medical field. RTI International's Responsive Virtual Human Technology has been applied to clinical examination skills training. Users conduct medical interviews of virtual humans displayed on a standard PC monitor. They interact with the virtual humans using a natural language interface combined with keyboard and mouse input [13]. RTI has also demonstrated preliminary validity of interactive virtual human experiences used to assess risky behavior in young adolescents [12] and to conduct informed consent interviews [11]. Also in the medical field, the Just VR system uses an immersive approach to have students experience and react to health emergency scenarios[23].

Most of the research in psychology deals with simulations where virtual humans are *spectators*. Pertaub et al. has treated fear of public speaking with virtual audiences [22]. They show that virtual audiences elicited the same behavior as real audiences from study participants. Recently virtual human technology has been utilized in treating post traumatic stress disorder for Iraq war veterans [20]. Bordnick revealed that a simulation of a *social* smoking experience using virtual humans yielding the highest cravings for participants over other non-social virtual experiences [4]. The company Virtually Better uses virtual humans in their psychological treatments[10].

Virtual humans have enjoyed the most success in entertainment. The movie industry has successfully used virtual humans as substitutes for real humans in

dangerous scenes or when the number of real humans required is prohibitive, such as in large battles. The characters are extremely life-like, but have completely scripted actions. Games are the most active users of virtual humans with nearly every new game utilizing them. Characters in games can be fully autonomous, but the interaction with them is generally unnatural, limited to pre-programmed commands and behaviors.

SYSTEM

The system layout is shown in Figure 2. The emphasis is on natural, transparent interaction with the virtual human, as though the user was interacting with a real human. User's can speak and gesture as though they were interviewing a real human. It is important to have natural interaction for direct comparison with real humans.

Speech is natural and unprompted. A wireless headset microphone is used for speech input. Speech recognition software, Dragon Naturally Speaking 8, translates the speech into text[19]. The translated text is displayed to the user to reduce user frustration when the system does not recognize speech accurately.

The head, hand, and torso motion of the user are tracked optically using passive infrared markers and two commodity, infrared sensitive, video cameras. The cameras are equipped with infrared lights and an infrared filter. This results in the infrared markers being the only visible objects in the video stream, making them easy to segment out. The segmented marker positions in each video stream are combined to produce the three dimensional position of each marker. Rigid clusters of markers and relative marker positions are used to register the marker positions to real world objects.

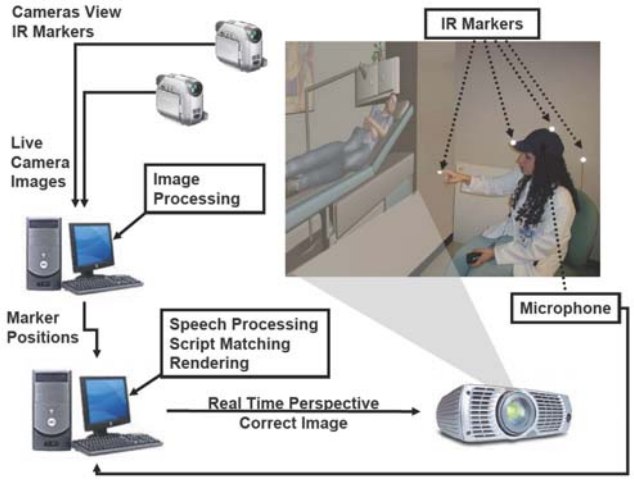

Figure 2. The System

The interaction is modeled as a question-answer session. The user asks a question, and then the system returns the appropriate speech and gesture. This question-answer pair is defined in a database created by medical faculty. Standardized patients are trained using a similar database. A lexical matching scheme is used to match the speech input to the database entry. The low frequency words, or keywords, are matched to the low frequency words in each database entry. Gesture input disambiguates questions such as "Does it hurt here?". In a recent analysis we conducted with the system, we found that, in most cases, over 60% of the student's input is matched to an appropriate response from the virtual human. Most of the errors (21%) were from questions we did not anticipate. The rest were speech recognition errors, variations in phrasing (such as negation), and other difficult to handle English wording. We find this sufficient because most students are able to complete the interview and come up with a diagnosis. A more thorough description our approach can be found in [7].

A patient's appearance directly affects the diagnostic process of the doctor. For a realistic appearing virtual humans, we use Haptek Corporation's full body virtual characters [8]. In addition to a realistic appearance, they have built in automatic animations including lip syncing, eye blinking, head following, and breathing. The voice for the virtual human is recorded audio of a standardized patient. Dickerson showed in [6] that recorded speech has advantages for natural virtual human interaction.

A projector is used as the display device. The choice of a projector over a head mounted display was made to leverage the surrounding environment. The surrounding environment is a real examination room. This maintains life-size proportions for the virtual human. A wall in the examination room serves as the projection surface. As done in CAVEs, the image is warped so that it appears correct from the tracked head position of the user [5]. Although it lacks a stereo effect, it provides most of the monocular depth cues. Monocular depth cues are important for judging size. The net effect is that the virtual world appears at life-size, as though it is an extension of the real examination room.

In the spectrum of virtual reality, augmented reality, mixed reality, and true reality, this system falls under the mixed reality category. While we use virtual reality techniques for rendering and interaction, the examination room the virtual character resides in is an extension of the real examination room. We believe this combination of real and virtual makes the VOSCE a far more immersive experience that if it were just a virtual reality system in a computer science laboratory.

STUDY

Design

The study uses a within subjects design depicted in Figure 3. Each participant performs both an Objective Structured Clinical Examination (OSCE) and a Virtual OSCE (VOSCE). The interview skills scores on the

	Scale	Description	Example(s)
Information	Number of 'yes' answers on 12 questions	Determines if critical information obtained from patient	*Description of Pain? Location of Pain?*
Process	Number of 'yes' answers on 13 questions	Determines interview performance	*Is there a logical pattern? Performs medical history?*
Quality	Average of 9 questions with scale (1 very poor, 2 poor, 3 good, 4 very good)	Determines interview quality	*Is empathetic? Displayed Appropriate Eye Contact? Body Lean?*
Overall	1,2,3-poor,4,5,6-adequate,7,8,9-good	The actual score assigned for the entire interview	*What is the overall score for this interaction?*

Table 1. The VOSCE/OSCE Interview Skills Checklist

VOSCE are then compared to the interview skills scores on the OSCE. These scores are given by a medical expert. Each participant also fills out a patient satisfaction survey, which rates the quality of the patient they interviewed.

The Essentials of Patient Care (EPC) class at the University of Florida educates students on medical interview skills. As part of the class, students perform OSCEs and are graded on their clinical skill. We integrated our study into the class during one of the OSCE evaluation sessions.

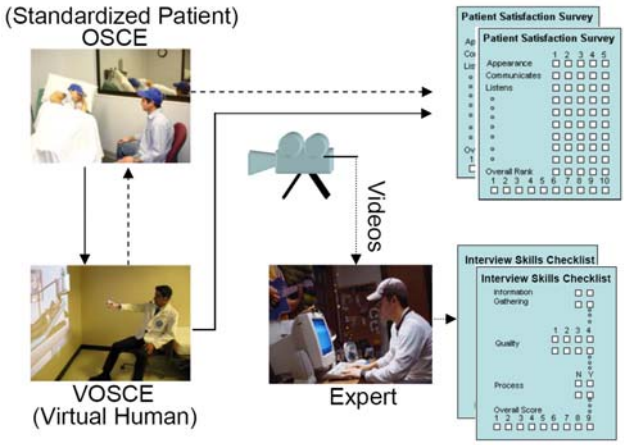

Figure 3. Study Design. Each Student is randomly assigned to the solid or dashed path. The student evaluates both the SP and the VH using the MaSP survey. An expert evaluates each student for both the VOSCE and the OSCE using the interview skills checklist.

The participants were randomly selected from the Fall 2005 EPC class. We had n=33 participants, 17 female and 16 male. They were all second year medical students and had the same experience interviewing standardized patients.

The scenario used for the VOSCE was a young Caucasian female with right lower quadrant (RLQ) pain. We used two different scenarios for the OSCE. One was a middle age Caucasian female complaining of chronic diarrhea (CHD) and the other a young Caucasian male complaining of indigestion (IND). These scenarios are typical for medical student training because standardized patients are available that can simulate them realistically.

The interactions took place at Harrell Center at the University of Florida, the standard testing center where University of Florida students perform OSCEs. This is important because it allowed the interviews to be conducted simultaneously, was familiar to the students, and is the place where a permanent VOSCE system could be installed.

Figure 4. Students begin their interview sessions by knocking on the door and entering the room. The second student from the left wearing the hat and microphone is interviewing the virtual human. The rest are interviewing standardized patients.

Procedure

Previous Pilot Studies

We ran pilot studies previously with the system, where the goal was to improve the system from a technical standpoint [14, 24]. Participants in these studies had no specific time constraints placed upon them during the procedure. The study procedure changed considerably in

this study in order to make the VOSCE and the OSCE run simultaneously under a strict time schedule. The process went from each student taking over 1 hour to complete to a maximum of 15 minutes. We had a maximum of 10 minutes that the participant could be in the examination room, so we removed the tutorial that taught the participant how to interact with the patient. While not having a practice session is uncommon in virtual reality research, a tutorial was largely unnecessary due to the natural interface.

Participant background gathering and speech training were performed a few days before the main study session. Surveys related to system satisfaction and presence were removed and there was no debriefing.

Current Study

In a typical OSCE scenario, a standardized patient awaits a medical student in each room of the testing center. During our study, however, Room 3 was occupied by the VH. All other rooms are occupied by the standardized patients. The students in the testing center were all part of the EPC course, but were not all part of the study. Only students who were seeing both the virtual human and a standardized patient with IND or CHD pain were included in the study.

All study participants are asked to sign an informed consent form and video release when they arrive at the testing facility. When it is their turn, one student stands outside each door and prepares for the interview by reading a chart describing the patient's vital statistics and chief complaint. The only difference for the student who interviews the VH is that they are outfitted with a hat used for tracking and a wireless microphone as shown in Figure 4.

The session begins with an audible signal saying "You may now start the station". The student at each room then enters and performs the required clinical examination. They are allowed 10 minutes. After 8 minutes a signal warns them to complete in 2 minutes. At the end of the session an audible signal says "Time is now up, please exit the station". Students are allowed to exit at any time during their interview. Once all students have completed their sessions, they are permitted to go back into the room. The patient then gives them feedback on their performance. The virtual human also has this capability. The participants then filled out the patient assessment questionnaire (discussed below) for each.

Metrics

Student Interview Skills Assessment

The standard way to evaluate students in OSCEs is a checklist. We used the interview skills assessment checklist. As seen in Table 1, 35 questions are divided into four separate areas: 9 questions on quality of interaction, 12 questions on amount of information gathered, 13 questions related to following proper interview process, and a single overall score for the interaction. The overall score is what actually determines pass or fail. The rest of the checklist exists as justification for a passing or failing mark.

This encounter was similar to my other patient encounters
I would use this as a practice tool
The patient communicates how he/she felt during the session
I can judge from the reactions of the patient whether she listens to me
I found this a worthwhile educational learning experience
The patient's appearance fits the role
The patient appears authentic
The patient is challenging/testing the student
The patient might be a real patient
What rank would you give the patient for this interaction? (1 lowest, to 10 highest)

Table 2. The Maastrict Assessment of Simulated Patients (MaSP). All questions except the rank use a 5 point likert scale – 1 (strongly disagree) 2 (disagree) 3 (neutral) 4 (agree) 5 (strongly agree)

Typically, either the SP is the evaluator right after the interview, or an expert performs the evaluation by watching video of the interview. Studies have shown that the SP can be as accurate or even more accurate than experts[16]. We used an expert because our VH does not have the capability to evaluate the student's interview skills reliably. This is a focus of future research, because in real world use, medical experts are expensive to use.

Patient Assessment

The Maastricht Assessment of Simulated Patients (MaSP) is a metric used to evaluate standardized patients[27]. It is a 5 point, likert-scale based survey administered to the medical student after an encounter with an SP. Similar to the interview skills checklist, the last question asks the medical student to rank the patient on a scale from one to ten. Sample survey items are shown in Table 2. We used the MaSP to evaluate both the SP and the VH. This enables us to study the correlation between the interview skills score obtained and the quality of the patient. This is important because the patient is the main difference between the VOSCE and the OSCE.

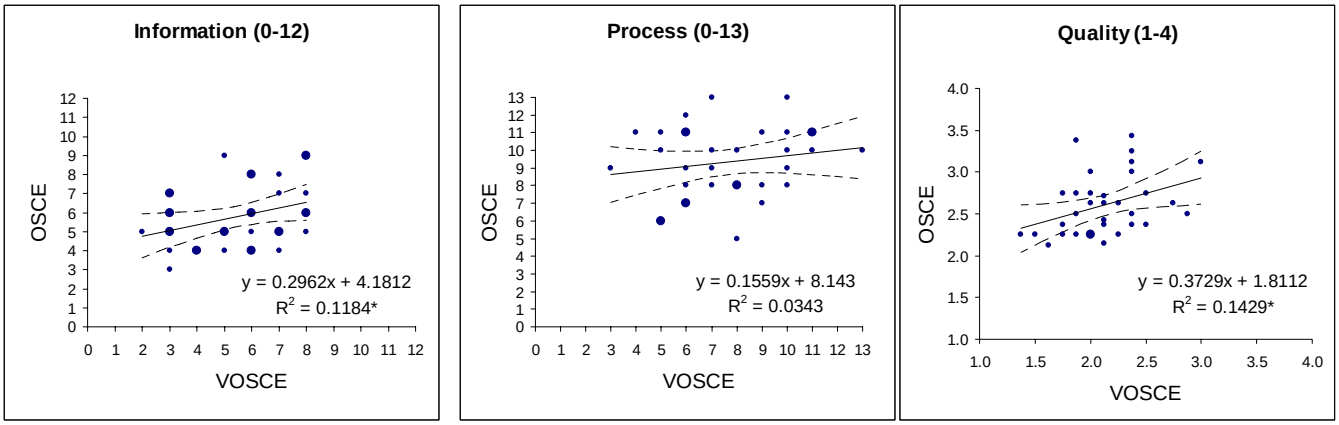

Figure 5. The interview skills checklist is broken up into three areas. The Process and Information scores indicate the number of 'yes' responses for each area. The Quality score is an average of the 9 questions related to quality of the interview. The dotted line represents the 95% confidence interval. The larger points indicate multiple students had those scores.

Experimental Validity

Internal

Internal validity is the truth of inferences made about cause and effect relationships in an experiment[17]. Overall we feel the experiment had strong internal validity. The claim is that performance is correlated between the VOSCE and OSCE because the same interview skills are required in both.

It is assumed that the interview skills of the participant are constant during the study. This may have been violated because the standardized patient and the virtual human both gave feedback after the session. This may have improved the participant's interview skills between interactions. While this improvement is a threat, it should not affect the correlation between the scores on the VOSCE and the OSCE, only shift the mean. The effect was also reduced by randomizing the order of the interactions.

The major threat to the internal validity of the study is from experimenter bias. Only one subject matter expert reviewed the video. While this could explain some of the correlation, we believe the student's interview skills are a more likely explanation. A solution to this problem is to use multiple experts, or observers trained by experts, to review video and measure their reliability.

We do not believe there were any other major confounding factors that affected the results. The procedure and metrics were held constant during the entire study and all participants completed the study.

External

External validity is the truth of generalizations about the real world made from experimental results[17]. We attempt to generalize our results to a situation where the VOSCE is used as a tool to evaluate medical students' interview skills; much like the how the OSCE is used currently.

The study participants were selected randomly from a class that all second year medical students are required to take. *The participants were not paid volunteers.* This is important because volunteers often are not representative of the true population and often are less critical of design flaws[17].

Participants tend to act differently when they know they are being tested[17]. The VOSCE is intended to be part of a testing environment. Thus, making the study appear to be part of the testing environment was important for establishing external validity. The VOSCE takes place inside a medical examination room, where clinical examinations will occur during their careers and where the OSCE takes place as well. During the study, participants were surrounded by other participants, their teachers, and their fellow students who were in the class but not part of the study. The effect of this on the participants was very noticeable. In contrast with pilot studies run on weekends with volunteer participants, participants acted more professionally, did not experiment as much with the system, and were more focused on the interview.

The major threat to external validity is that this study was run with students from just one medical education institution. There is no guarantee that this will translate well to students of other medical schools. Further, the students were beginning 2nd year students. More work needs to be done to show that regardless of educational level, performance is correlated between virtual humans and standardized patients.

RESULTS AND DISCUSSION

All students completed the study, came up with a diagnosis, and were able to do so in the time allotted to

them. All of this occurred concurrently with the OSCE in the normal training environment. While this is critical for direct comparison, it is also encouraging from a feasibility standpoint. It shows that the VOSCE can integrate into the existing infrastructure of clinical skills education.

Interview Skills Checklist

We analyze the results from the interview skills checklist by category, comparing group mean differences through an ANOVA, and correlation through regression analysis.

The last question on the checklist asked the evaluator to judge the overall quality of the interview on a scale from 1 to 9, 1 being the worst, 9 being the best. This is the grade that a student would get for the interview. A graph of same student overall performance in the VOSCE and the OSCE is shown in Figure 6. There is a significant ($p<.005$) correlation in the overall rating of same student VOSCE to OSCE interactions. The Pearson correlation coefficient was $r=.497$ ($r^2=.247$).

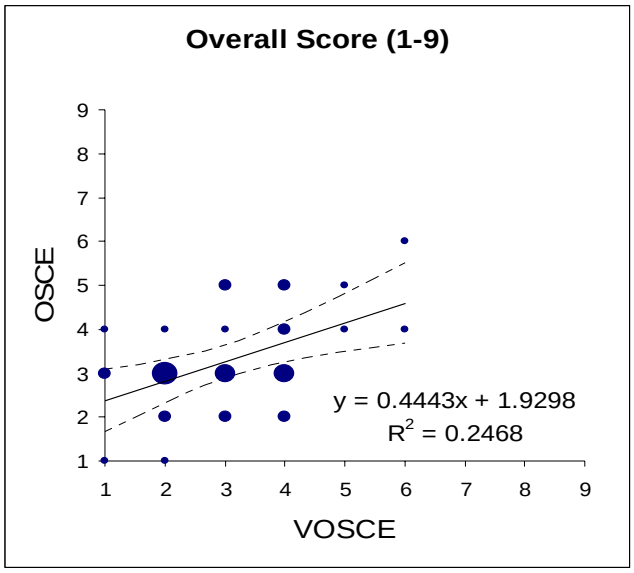

Figure 6. There is a significant ($p<.005$) correlation in the overall score for VOSCE and OSCE sessions. The dotted lines indicate the 95% confidence interval. Larger points indicate multiple students had those scores.

The worst students performed the worst in both the VOSCE and the OSCE and the best students performed the best in both the VOSCE and the OSCE. The study was run with early second year students with equivalent training for clinical examinations. At this stage in their education, medical students are just beginning to practice their clinical examination skills. Both the VOSCE and the OSCE are able to indicate the maturity of the student. No student scored above a 6 for either the VOSCE or the OSCE and only a few scored above a 4. In its current state, we feel the VOSCE can be used right away by students needing additional practice. Standardized patients can only be used during pre-organized training sessions. The VOSCE can be available anytime. The standardized patient training is still necessary, however, for training students for scenarios requiring complex behaviors that a virtual human can not yet simulate.

While the overall score determines the student's grade for the interaction, the medical expert also evaluated the students in three subcategories, Information, Process, and Quality. Figure 5 shows graphs of each student's performance in both the VOSCE and OSCE by category. The information and quality areas showed significant, although small correlation. The null hypothesis could not be rejected for the process area. While this did not seem to factor much into the overall score, it is clearly an area in which the VOSCE could improve.

A multivariate ANOVA was conducted on the interview skills data for the four categories: information, process, quality and overall score. There were two factors. The first factor was interaction type, VOSCE or OSCE. The second factor was gender, which has been shown to have a large effect in studies with virtual characters [28]. There was not a multivariate effect found for gender or interaction type×gender. There was a significant ($p<.001$) multivariate effect found for interaction type. A univariate analysis showed a significant effect of interaction type on both the quality ($p<.001$) and the process ($p<.01$) areas. The mean scores for each area are compared in Figure 7.

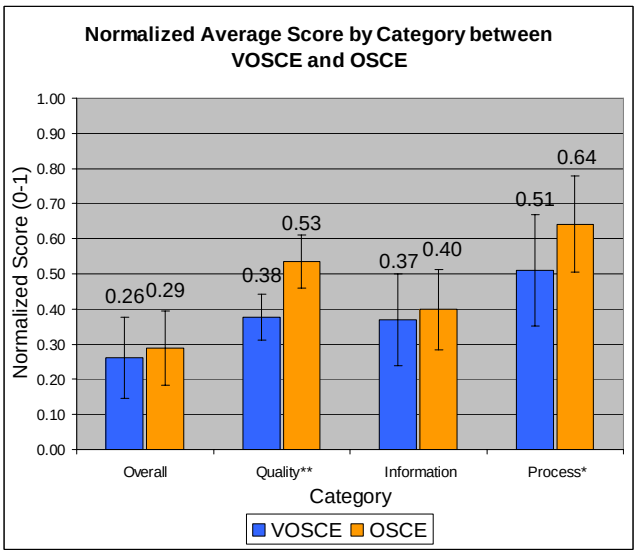

Figure 7. Scores for each subject area of the interview skills checklist are normalized between 0 and 1. **p<.01,*p<.05

The overall score was not significantly different between the VOSCE and the OSCE. The amount of information obtained was also not different. Also, the correlation in these areas was significant. This means that, for testing

overall skill and information gathering skill, the VOSCE could be used in place of the OSCE.

The quality ($p<.01$) and process ($p<.05$) scores showed significant difference between scenario types. These are areas where the VOSCE suffers from technical limitations. The individual questions in the quality area of the interview skills checklist were mostly related to non-verbal behavior such as body lean, head nod, and eye contact. Maintaining appropriate non-verbal behavior might have been difficult when we required that the student wear a baseball cap for tracking and a microphone for speech input. In addition, their motion was hindered because of occlusion of the projected image and visual field of the tracking cameras. Expectations of the virtual human's ability to respond to appropriate non-verbal behavior may have also been lower. Regardless, there was still a significant correlation in quality ($r=.378$, $r^2=.143$, $p<.05$).

The process, quality, and information scores are supposed to be reflected in the overall score. The overall scores were not different, but there were significant differences in the process and quality scores. This is explained by analyzing the interview skills scores. While the interview skills checklist has a high degree of internal consistency (Cronbach's $\alpha=.78$), the overall score had the highest correlation with the information score ($r=.77$, $p<.001$).

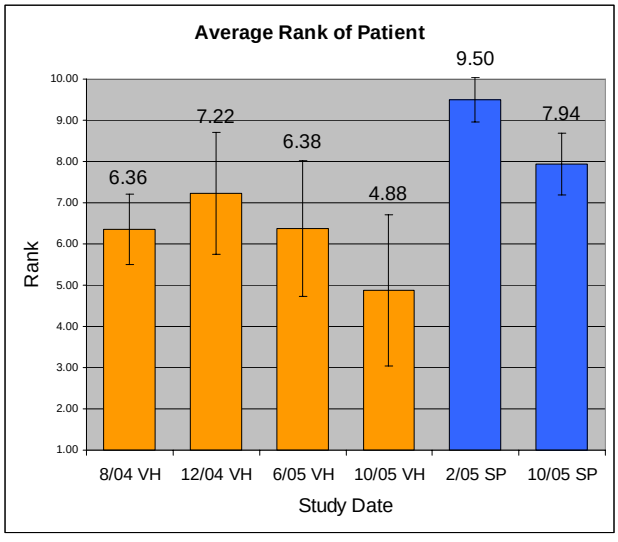

Figure 8. The overall rank given to the patient by participants in the MaSP survey separated by type of patient and study date. The drop from previous studies to the current October 2005 study is significant at the p<.001 level.

MaSP

While the interview skills checklist looks at how good the medical student is at being a doctor, the MaSP looks at how good the virtual human or standardized patient is at being a patient. Participants filled out the MaSP after their VOSCE and OSCE. The wrong scale was used for day 1 of the study. This error resulted in MaSP data for 16 participants to be excluded.

Our analysis of the MaSP results showed a large drop in participant satisfaction for both the SP ($p<.001$) and VH ($p<.05$) groups relative to previous study results[14, 15, 24]. In fact, while previous studies have been shown the virtual human to be close to the national average for standardized patients (7.2), the mean score for the virtual human in this study is only 4.88(σ =1.83) This is illustrated in Figure 8.

We believe the main reason for this drop is volunteer bias. Volunteers were not used in this study, whereas only paid volunteers were used in previous studies. We do not believe that this is a result of system changes because both the virtual human and the standardized patient ranks were affected similarly.

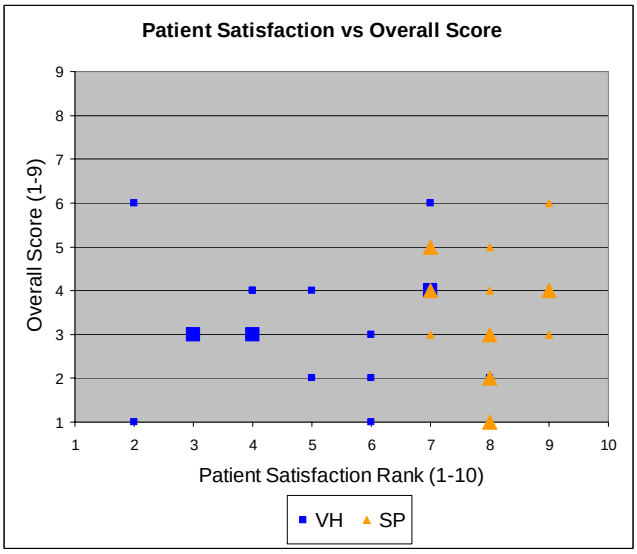

Figure 9. Patient Satisfaction as determined by the rank given by the participant is compared to the participants overall score as determined by the evaluator. Larger data points indicate multiple students had those scores.

Performance of the patient was compared to performance of the participant. No observable correlation was found between patient rank on the MaSP and overall score on the interview skills checklist as seen in Figure 9. This is concerning, because we operate under the assumption that the quality of the virtual human is important to validity of the VOSCE. This does not seem to be shown by the data. Overall performance is most likely not a linear function of patient satisfaction. *We hypothesize that the fidelity of the virtual human as a patient must be good enough so the student can demonstrate good interview skills and perform a complete interview.* For example, a virtual human that could not respond to any input would have made completing the interview frustrating or impossible.

We need to determine at what threshold of fidelity a virtual human experience becomes successful.

CONCLUSIONS AND FUTURE WORK

The purpose of this work was to validate our virtual human experience for clinical examination interview skills testing. The validation tests if medical interview skills used during an interview with a virtual human are the same as those used during an interview with a real human.

Interview skills performance was determined by a checklist containing the required elements for a clinical examination interview. This checklist was filled out by a medical expert watching video of the interview. Participants conducted the interview portion of a clinical examination with a virtual human and then a standardized patient. The correlation in performance was then computed. *We found a medium correlation ($r(33)=.49$, $p<.005$) in overall performance*.

We found that *participants ranked both the virtual human and standardized patient much lower than in previous studies.* Participants in previous studies were paid to participate and the interactions took place on a weekend, outside of normal classroom activities. Students in the current study were assigned as part of a class requirement. While the ranking was lower for the virtual human, *we did not find a significant correlation between ranking and performance for either the virtual human or the standardized patient*.

Overall, *the study had a strong internal and external validity.* While many virtual reality studies have strong internal validity, few have strong external validity. Integrating the VOSCE into the existing infrastructure for medical student clinical examination education gives it a high level of external validity. The VOSCE was run with *real users* in the *real environment*, *and it worked*.

The VOSCE is expanding as an evaluation and training tool. Studies are planned to look at the VOSCE for anxiety reduction when taking sexual histories, and for training students on racial issues. This is something that is infeasible, or impossible to do using trained actors. A pelvic simulator is being combined with the virtual human technology from VOSCE. Students will perform a real physical examination while getting instruction from a virtual human. Visualization software is being created that will allow a student to review their performance in ways never before possible. With the current video review mechanism students only get one or two perspectives on the interview. With the new visualization software they will be able to view their interview from any perspective in the room – even from the patient's.

Other work will look into how system factors affect performance. The current technology limits how the student can interact with the virtual human. A new study will determine how display device affects performance. There may be a difference in how a person interacts with a virtual human if the virtual human is displayed on an ordinary monitor, a projector, or a head mounted display.

This work shows that a virtual human experience can be as effective as a real human experience in real world interpersonal skills education. We believe that the natural interface provided by speech and gesture recognition combined with the choice of a life-size presentation of the virtual human is what makes this possible. We hope that this work encourages the development and study of new virtual human applications.

OTHER AUTHORS AND ACKNOWLEDGEMENTS

We thank Robert Dickerson for his help running the user study. We thank Dr. Marc Cohen and Dr. Juan Cendan for help analyzing the data. We also thank Dr. Rebecca Pauly for allowing us to use students from her Fall 2005 EPC class. Finally, we thank Dr. Margaret Duerson for providing support at the Harrell Center. This work was partially supported by a UF Alumni Fellowship and a grant from the UF College of Medicine.

REFERENCES

1. Alessi, N.E. and Huang, M.P., Evolution of the Virtual Human: From Term to Potential Application in Psychiatry. *CyberPsychology & Behavior* 3 (2000), 321-326.
2. Badler, N.I., Erignac, C.A. and Liu, Y., Virtual Humans for Validating Maintenance Proceedures. *Communications of the ACM* 45 (2002), 56-63.
3. Badler, N.I., Phillips, C.B. and Webber, B.L. *Simulating Humans: Computer Graphics, Animation, and Control*. Oxford University Press, New York, 1993.
4. Bordnick, P.S., Graap, K.M., Copp, H.L., Brooks, J. and Ferrer, M., Virtual Reality Cue Reactivity Assessment in Cigarette Smokers. *Cyberpsychology & Behavior* 8,5 (2005), 487-492.
5. C. Cruz-Neira, Sandin, D.J. and DeFanti, T. The design and implementation of the CAVE. In *SIGGRAPH '93*, (1993),135-142.
6. Dickerson, R., Johnsen, K., Raij, A., Lok, B., Bernard, T., Stevens, A. and Lind, D.S. Virtual Patients: Assessment of Synthesized Versus Recorded Speech. In *Proceedings of Medicine Meets Virtual Reality*, (2005),114-119.
7. Dickerson, R., Johnsen, K., Raij, A., Lok, B., Hernandez, J., Stevens, A. and Lind, D.S. Evaluating a Script-Based Approach to Simulating Patient-Doctor Interaction. In *Proceedings of SCS 2005 International Conference on Human-Computer Interface Advances for Modeling and Simulating*, (2005),79-84.

8. Full Body Characters. www.haptek.com.
9. Hill, R., Gratch, J., Marsella, S., Rickel, J., Swartout, W. and Traum, D., Virtual Humans in the Mission Rehearsal Exercise System. *Künstliche Intelligenz* 4,03 (2003), 5-10.
10. Hodges, L.F., Anderson, P., Burdea, G., Hoffman, H. and Rothbaum, B., Treating Psychological and Physical Disorders with VR. *IEEE Computer Graphics and Applications* (2001), 25-32.
11. Hubal, R.C. and Day, R.S., Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. *Journal of Biomedical Informatics* (2006).
12. Hubal, R.C., Fishbein, D.H. and Paschall, M.J. Lessons Learned using Responsive Virtual Humans for Assessing Interaction Skills. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference(I/ITSEC) 2004*, (2004).
13. Hubal, R.C., Kizakevich, P.N., Guinn, C.I., Merino, K.D. and West, S.L., The Virtual Standardized Patient: Simulated Patient-Practitioner Dialog for Patient Interview Training. *Studies in Health Technology and Informatics* 70 (2000), 133-138.
14. Johnsen, K., Dickerson, R., Raij, A., Harrison, C., Lok, B., Stevens, A. and Lind, D.S., Evolving an Immersive Medical Communication Skills Trainer. *Presence: Teleoperators and Virtual Environments* 15,1 (2006), 33-46.
15. Johnsen, K., Dickerson, R., Raij, A., Lok, B., Jackson, J., Shin, M., Hernandez, J., Stevens, A. and Lind, D.S. Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills. In *Proceedings of the 2005 IEEE Conference on Virtual Reality*, (2005),179-186, 324.
16. McLaughlin, K., Gregor, L., Jones, A. and Coderre, S., Can standardized patients replace physicians as OSCE examiners? *BMC Medical Education* 6,1 (2006), 12.
17. Mitchell, M. and Jolley, J. *Research Design Explained*. Harcourt, New York, NY, USA, 2001.
18. Nass, C. and Moon, Y., Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56 (2000), 81-103.
19. Dragon Naturally Speaking 8 Professional. www.nuance.com.
20. Pair, J., Allen, B., Dautricourt, M., Treskunov, A., Liewer, M., Graap, K., Reger, G. and Rizzo, A. A virtual Reality Exposure Therapy Application for Iraq War Post Tramatic Stress Disorder. In *Proceedings of 2006 IEEE Conference on Virtual Reality*, (2002),67-72.
21. Park, R.S., Chibnall, J.T., Blaskiewicz, R.J., Furman, G.E., Powell, J.K. and Mohr, C.J., Contruct Validity of an Objective Structured Clinical Examination (OSCE) in Psychiatry: Associations with the Clinical Skills Examination and Other Indicators. *Academic Psychiatry* 28,2 (2004), 122-128.
22. Pertaub, D., Slater, M. and Barker, C., An experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence: Teleoperators and Virtual Environments* 11 (2002), 68-78.
23. Ponder, M., Herbelin, B., Molet, T., Scherteneib, S., Ulicny, B., Papagiannakis, G., Magnenat-Thalmann, N. and Thalmann, D. Interactive Scenario Immersion: Health Emergency Decision Training in the JUST Project,. In *VRMHR 2002*, (2002).
24. Raij, A., Johnsen, K., Dickerson, R., Lok, B., Cohen, M., Bernard, T., Oxendine, C., Wagner, P. and Lind, D.S. Interpersonal Scenarios: Virtual approx Real? In *Proceedings of the 2006 IEEE Conference on Virtual Reality*, (2006),80-88, 378.
25. Thórisson, K.R. and Cassell, J. Why Put an Agent in a Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue. In *Proceedings of Lifelike Computer Characters '96*, (1996),44-45.
26. Vinayagamoorthy, V., Steed, A. and Slater, M. Building Characters: Lessons Drawn from Virtual Environments. In *Proceedings of Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, (2005),119–126.
27. Wind, L.A., Dalen, J.v., Muijtjens, A.M.M. and Rethans, J.-J., Assessing Simulated Patients in an Educational Setting: the MaSP (Maastricht Assessment of Simulated Patients). *Medical Education* 38,1 (2004), 39-44.
28. Zanbaka, C., Goolkasian, P. and Hodges, L. Can a Virtual Cat Persuade You?: The Role of Gender and Realism in Speaker Persuasiveness. In *CHI 2006*, ACM Press(2006),1153 - 1162.