Evaluation of Justina: A Virtual Patient with PTSD

Patrick Kenny, Thomas D. Parsons, Jonathan Gratch, and Albert A. Rizzo

Institute for Creative Technologies,
University of Southern California
13274 Fiji Way Marina Del Rey, CA 90292, USA
{kenny,tparsons,gratch,rizzo}@ict.usc.edu

Abstract. Recent research has established the potential for virtual characters to act as virtual standardized patients VP for the assessment and training of novice clinicians. We hypothesize that the responses of a VP simulating Post Traumatic Stress Disorder (PTSD) in an adolescent female could elicit a number of diagnostic mental health specific questions (from novice clinicians) that are necessary for differential diagnosis of the condition. Composites were developed to reflect the relation between novice clinician questions and VP responses. The primary goal in this study was evaluative: can a VP generate responses that elicit user questions relevant for PTSD categorization? A secondary goal was to investigate the impact of psychological variables upon the resulting VP Question/Response composites and the overall believability of the system.

Keywords: Virtual Humans, Virtual Patients, Psychopathology.

1 Introduction and Background

The potential of using virtual humans as virtual standardized patients (VP) for use in clinical assessments, interviewing and diagnosis training is becoming recognized as the technology advances [2,3]. These VPs are embodied interactive agents [4,6,11,24,26] who are designed to simulate a particular clinical presentation of a patient with a high degree of consistency and realism [15, 28]. VPs have commonly been used to teach bedside competencies of bioethics, basic patient communication, interactive conversations, history taking, and clinical decision making. [5,18,26] VPs can provide valid, reliable, and applicable representations of live patients [33]. Research into the use of VPs in psychotherapy training is in its nascent stages [8,14,20]. Since virtual humans and virtual environments can allow for precise presentation and control of dynamic perceptual stimuli (visual, auditory, olfactory, gustatory, ambulatory, and haptic conditions), conversations and interactions, they can provide ecologically valid assessments that combine the control and rigor of laboratory measures with a verisimilitude that reflects real life situations [1,18,23,30]. Although progress has been made toward establishing systems that are sensitive to component psychological processes, more studies are required to understand the effectiveness of these systems for training and education, to measure the believability of the characters with respect to their verbal and non-verbal behavior and how different genders, races and personality or interview styles interact with the characters.

H. Prendinger, J. Lester, and M. Ishizuka (Eds.): IVA 2008, LNAI 5208, pp. 394–408, 2008.

This current project builds on our previous work in building a VP for conduct disorder [15] and aims to improve child and adolescent psychiatry residents, and medical students' interview skills and diagnostic acumen for a difficult subject through practice with a female adolescent virtual human with post-traumatic stress disorder (PTSD). This interaction with a VP provides a context where immediate feedback can be provided regarding trainees' interviewing skills in terms of psychiatric knowledge, sensitivity, and effectiveness. Use of an embodied natural language-capable virtual character is beneficial in providing trainees with exposure to psychiatric diagnoses such as PTSD that is prevalent in their live patient populations and believed to be under-diagnosed due to difficulty in eliciting pertinent information. Virtual reality patient paradigms, therefore, will provide a unique and important format in which to teach and refine trainees' interview skills and psychiatric knowledge.

In this paper we describe a series of subject tests of a virtual patient system performed with medical students to evaluate its usefulness and effectiveness as a medium to communicate with the students. The evaluation consisted of an assessment of the system as a whole through questionnaires and data collection of the questions and responses in the interview. Additionally we investigate the relationship of the questions with a number of psychological variables such as openness to interaction with the VP and willingness to be immersed in the virtual environment.

2 Designing a Patient with Post Traumatic Stress Disorder (PTSD)

2.1 Virtual Justina

One of the challenges of building complex interactive VPs that can act as simulated patients has been in enabling the characters to act and carry on a dialog like a real patient that has the specific mental condition in the domain of interest. Additional issues involve the breadth and depth of expertise required in the psychological domain to generate the relevant material for the character and dialog. In our first attempt to design a VP 'Justin'[15] we choose a domain, conduct disorder, that was more forgiving of inappropriate responses to user questions and where the patient would be somewhat resistant to answering questions. Inappropriate or out of domain responses

Fig. 1. Justina Virtual Patient

were seen as part of the disorder and this did not negatively impact the interview process. The current domain of PTSD is less forgiving and requires the system to respond appropriately based on certain criteria for PTSD as described in the Diagnostic and Statistical Manual of mental disorders (DSM-IV) category (309.81) [9]. For the PTSD domain we built an adolescent girl character called Justina, see Figure 1. Justina has been the victim of an assault and shows signs of PTSD. The technology used for the system is based on the virtual human technology developed at USC [16,29] and is the same as what was used with the previous VP 'Justin'. The system uses speech recognition, question / response and a procedural animation system to control the character.

2.2 PTSD Domain

The experience of victimization is a relatively common occurrence for both adolescents and adults. However, victimization is more widespread among adolescents, and its relationship to various problem outcomes tends to be stronger among adolescent victims than adult victims. Whilst much of the early research on the psychological sequelae of victimization focused on general distress or fear rather than specific symptoms of PTSD, anxiety, or depression, studies have consistently found significant positive correlations between PTSD and sexual assault, and victimization in general and violent victimization in particular [22]. Although there are a number of perspectives on what constitutes trauma exposure in children and adolescents, there is a general consensus amongst clinicians and researchers that this is a substantial social problem [25]. The effects of trauma exposure manifest themselves in a wide range of symptoms: anxiety, post-trauma stress, fear, and various behavior problems. New clinicians need to come up to speed on how to interact, diagnose and treat this trauma.

According to the most recent revision to the American Psychiatric Association's DSM Disorders, PTSD is divided into six major categories; refer to the DSM-IV category 309.81 [9] for a full description and subcategories;

A. Past experience of a traumatic event and the response to the event.
B. Re-experiencing of the event with dreams, flashbacks and exposure to cues.
C. Persistent avoidance of trauma-related stimuli: thoughts, feelings, activities or places, and general numbing such as low affect and no sense of a future.
D. Persistent symptoms of anxiety or increased arousal such as hyper vigilance or jumpy, irritability, sleep difficulties or can't concentrate.
E. Duration of the disturbance, how long have they been experiencing this.
F. Effects on their life such as clinically significant distress or impairment in social or educational functioning or changes in mental states.

Diagnostic criteria for PTSD includes a history of exposure to a traumatic event in category A and meeting two criteria and symptoms from each B, C, and D. The duration of E is usually greater than one month and the effects on F can vary based on severity of the trauma. Effective interviewing skills are a core competency for the clinicians, residents and developing psychotherapists who will be working with children and adolescents exposed to trauma. A clinician needs to ask questions in each of these categories to properly assess the patient's condition.

2.3 Question / Response Categorization

Domain building for the VP consisted of role-playing sessions to gather the verbal and non-verbal behavior for the patient along with the set of questions typically asked by a clinician. Additionally, iterative discussions with psychiatry faculty from the Keck School of medicine at USC were performed to enhance the corpus of questions and responses. The goal was to build enough of the domain to cover the six categories in the PTSD DSM criteria and cover the kinds of questions people would ask a patient. The corpus was used for the statistically natural language question/response system [17, 32]. The natural language system works by selecting responses based on input questions. The set of questions and responses are manually mapped by a domain expert. For this application domain there were a total of 459 questions that mapped roughly 4 to 1 to a set of 116 responses. The aim was to build the domain corpus with what we could anticipate and then elicit questions from the user that s/he may ask of the VP for the specific traumatic experience and use those questions in an iterative process to further build the corpus. Since PTSD falls in the diagnostic category of anxiety disorders, rather than assessing for all of the specific criteria, we initially focused at a high level upon the six major clusters of symptoms following a traumatic event. While this did not give the character depth but breadth, for initial testing this seemed prudent. Next, we developed two additional categories that we felt would aid in assessing user questions and VP responses that are not included in the DSM;

G. A general category meant to cover questions regarding establishing rapport, establishing relations, clarifications, opening and closing dialog.
H. A category to cover accidental mouse presses or miscellaneous items. Users interact with the system with speech, however they need to press the mouse button while talking. Sometimes when people are thinking they have a tendency to press the button then release without saying anything, this causes the system to respond with an off topic response, and can confuse new users.

Table 1 is an example of some questions and responses from Justina for each of the six categories. Once all of the responses were established a voice actor was used to record the voice for Justina to be used by the system.

Table 1. Question / Response Categorization

Category	User Question	Justina Response
1(A) Trauma	So, what happened to you that night?	Something really bad happened.
2(B) Re-experience	Do you still think about what happened?	Sometimes I feel like the attack is happening all over again
3(C) Avoidance	Do you go out with your friends?	I just stay away from everyone now.
4(D) Arousal	Do you feel jumpy?	I feel like I have to watch my back all the time.
5(E) Duration	How long has this been going on?	A few months
6(F) Life Effect	Are you upset?	Sometimes I don't do anything but stay in my room and cry.
7(G) Communication	Hi Justina, I'm Doctor..	Hello
8(H) Other	Button Press	I don't get what you mean.

3 Method

Although our primary goal in this study was evaluative: to assess the effectiveness of a virtual standardized patient to generate responses that elicit user questions relevant for a virtual character that has PTSD, a secondary goal was to investigate the relationship between a number of psychological variables and the resulting VP Question/Response composite. An important issue in the study of intelligent virtual agents is to identify under what circumstances a person interacting with the virtual agent is open to the interaction. We were interested in the psychological variables of hypnotizability and absorption [31] as well as immersiveness [37] and presence [10,21] in relation to a person's experience of an interaction with the VP. Although these variables have been little explored, results from a recent study reveal that physiological arousal appeared to be moderated by participant hypnotizability and absorption levels [19,36]. High-absorption individuals may be more capable of imagining that the VP has PTSD when it is suggested. It was hypothesized that participants' scores on measures of absorption, immersion, and presence would be positively and significantly correlated with a measure of their VP Question/Response composite.

3.1 Participants

Participants were asked to take part in a study of novice clinicians interacting with a VP system. They were not told what kind of condition the VP had if any. Two recruitment methods were used: poster advertisements on the university medical campus; and email advertisement and classroom recruitment to students and staff. A total of 15 people (6 females, 9 males; mean age = 29.80, SD 3.67) took part in the study. Ethnicity distribution was as follows: Caucasian = 67%; Indian = 13%; and Asian = 20%. The subject pool was made up of three groups: 1) Medical students (N=7); 2) Psychiatry Residents (N=4); 3) Psychiatry Fellows (N=4). For participation in the study, students were able to forgo certain medical round time with the time spent in the interview and questionnaires, which took approximately 45 minutes.

3.2 Setup

The VP system consisted of the virtual character Justina, as seen in Figure 1 and 2, along with a headset for speech input and mouse button which was required to be

Fig. 2. Testing setup and interaction

pressed while speaking. A control station was adjacent to the subject to run the system and log the data. Cameras were setup to record the subjects face and the interaction with the VP from the side for later post processing analysis and review.

3.3 Process and Procedure

Medical students currently perform interview training with human actors acting as standardized patients. The actors portray some clinical problem in what is called an Objective Structured Clinical Examination (OSCE) [13,34,35]. These tests typically take from 20-30 minutes, a faculty member watches the student perform and students are videotaped. The evaluation consists of self assessment rating along with faculty assessment and a review of the videotape. This practice is common, although varies based on the actors, available faculty members and space and time at the university. Although schools commonly make use of standardized patients to teach interview skills, the diversity of the scenarios that standardized patients can characterize is limited by the availability of human actors and their skills at portraying the condition. Additionally the actors most likely vary their performance from subject to subject and location to location. This is an even greater problem when the actor needs to be an adolescent, elder or portray a difficult condition. Our process is similar to an OSCE, but the actor is replaced with a virtual patient and an observer is replaced by video recording. Using virtual patients will allow standard performances for all subjects.

The subject testing was divided into three phases, a pre-test and pre-questionnaire, the interview and a post-questionnaire. The pre-questionnaire was performed in a separate room from the interview and took about 10 minutes. For the interview the participants were asked to perform a 15 minute interaction with the VP and assess any history or initial diagnosis of a condition of the character. The participants were asked to talk normally as they would to a standardized patient, but were informed that the system uses speech recognition and was a research prototype. They were free to ask any kind of question and the system would try to respond appropriately. At the end of the 15 minute exchange they would be sent to another room to take the post-questionnaire. Data was logged during the interview for later processing. The video recordings and system logs of the interaction could be re-played for review, critique and commentary by child and adolescent psychiatry attendings, as a teaching tool for residents, or for groups of medical students learning about PTSD, or even for students using distance learning that don't have access to this technology.

The data in the system was logged from various modules. Figure 2 is a diagram of how the user interacts with the VP system and the logging and annotation pipeline. First the user speech is recorded from the automated speech recognition (ASR) engine and later transcribed, this is the actual text said by the user. Next the output text from the ASR is logged, this text is usually not 100% accurate due to speech models, accents, or voice types. This output when compared with the actual text will give the accuracy of the speech engine, due to time constraints we were not able to compute the accuracy of the ASR, but know that it could be improved. The ASR output is sent to the natural language (NL) statistical question/response system. The NL system records a transcript of the entire dialog session, this is used later to help analyze the interaction. System messages are logged and can be used to reconstruct what was happening in all parts of the system as needed. Cameras record participant's facial expressions and system interaction with the patient for analysis at a later time.

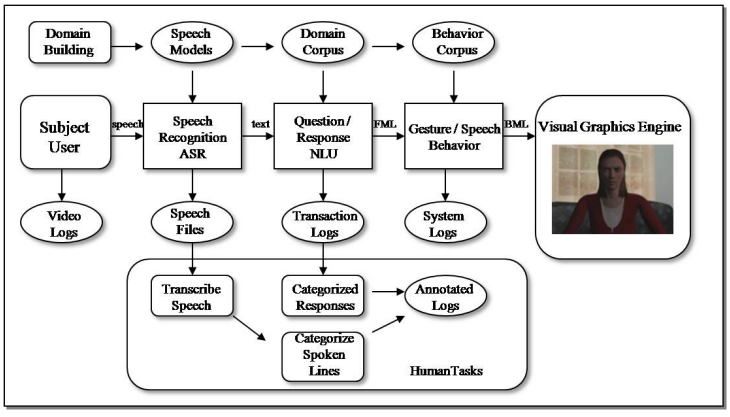

Fig. 3. Interaction and Data Logging Pipeline

3.4 Measures

As mentioned above, an important issue in the study of intelligent virtual agents is to identify under what circumstances a person interacting with the agent is open to the interaction. By this we mean how open is the person interacting with the VP open to new experiences and interacting with novel technologies. Again, results from a recent study that we conducted reveal that physiological arousal appeared to be moderated by participant openness to such interactions [19,36]. High-absorption individuals may be more capable of imagining that the VP has PTSD when it is suggested.

The following standardized and unstandardized measures were used to assess the impact of absorption and immersiveness upon the "believability" of the system. Prior to the experiment itself, the subjects were required to fill in the following standardized questionnaires: 1) Tellegen Absorption Scale (TAS). The TAS questionnaire aims to measure the subject's openness to absorbing and self-altering experiences. The TAS is a 34-item measure of absorption, and is a widely used questionnaire with well established reliability and validity [31]. 2) Immersive tendencies questionnaire (ITQ). The ITQ measure individual differences in the tendencies of persons to experience "presence" in an immersive VE. The majority of the items relate to a person's involvement in common activities. While some items measure immersive tendencies directly, others assess respondents' current fitness or alertness, and others emphasize the user's ability to focus or redirect his or her attention. The ITQ is comprised of 18 items, and each is rated on a 7-point scale and is a widely used questionnaire with well established reliability and validity [37]. Subjects also completed unstandardized measures that were developed specifically for this protocol: 1) Virtual Patient Pre-Questionnaire (VPQ1). This scale was developed to establish basic competence for interaction with a virtual character that is intended to be presented as one with PTSD, although no mention of PTSD is on the test. 2) Justina Pre-questionnaire (JPQ1). We developed this scale to gather basic demographics and ask questions related to the user's openness to the environment and virtual reality user's perception of the technology and how well they think the performance will be. There were 5 questions regarding the technology and how well they thought they might perform with the agent.

After the experiment the subjects were instructed to fill in the following standardized questionnaire: 1) Presence questionnaire (PQ). The Presence Questionnaire is a common measure of presence in immersive virtual reality. Presence has been described of as comprising three particular characteristics: sense of being within the VE; extent that the VE becomes the dominant reality for users; and extent to which users view the VE as a place they experienced rather than simply images they observed. The PQ is a widely used questionnaire with well established reliability and validity [37] Subjects were also asked to complete unstandardized questionnaires developed specifically for this protocol: 1) Justina Post-questionnaire (JPQ2). We developed this scale to survey the user's perceptions related to their experience of the virtual environment in general and experience interacting with the virtual character in particular the patient in terms of its condition, verbal and non-verbal behavior and how well the system understood them and if they could express what they wanted to the patient. Additionally there were questions on the interaction and if they found it frustrating or satisfying. There were 25 questions for this form. 2) Virtual Patient Post-questionnaire (VPQ2). This scale was exactly the same as the Virtual Patient Pre-questionnaire and will be used in the future for norming of a pre-post assessment of learning across multiple interactions with the VP. In the future we will also include social presence and rapport scales and include a control set that will just go thru a fixed script with the interview.

3.5 Data Analytics

Participants completed the VPQ1; JPQ1; TAS; and ITQ prior to the VP trial. Following this, participants received instructions on how to interact with the patient, then the trial started. After 15 minutes the trial was completed, and participants then completed a PQ; JPQ2; and VPQ2. When all the trials were completed the speech for each participant was transcribed and annotated with one of the categories in Table 1.

Here we focused on effective interview skills—a core competency for psychiatry residents and developing psychotherapists. The keys aspects of the interview that we looked at were: interpersonal interaction; attention to the VP's vocal communications, as well as verbal and non-verbal behavior. Specifically, we wanted to assess whether the clinician established and maintained rapport, as well as ask questions related to the reason for referral. We also wanted to assess whether the user (clinician in training) made attempts to gather information about the VP's problems. Finally, we wanted to see if the user would attempt detailed inquiry to gain specific and detailed information from the VP, separating relevant from irrelevant information.

VP Question/Response Composite

Question/response composites were developed to reflect the shared variance existing between the responses of a VP simulating PTSD in an adolescent female and of DSM IV TR-specific Questions (from novice clinicians) that are necessary for differential diagnosis. The question/response composites drawn from novice clinician questions and VP responses were referred to as VP Question/Response composites or (VP_QR'). Again, the primary goal in this study was evaluative and the VP_QR' scores were calculated to assess whether a virtual standardized patient could generate responses that elicit user questions relevant for PTSD categorization. For the VP_QR'

scores, we first calculated eigenvalues via least squares procedures and separate composite measures were created for each observation. The resulting weights were used in conjunction with the original variable values to calculate each observation's score. The VP_QR' scores were standardized according to a z-score.

Primary and Secondary Analyses

To assess whether the responses of a VP simulating PTSD in an adolescent female could elicit a number of DSM IV TR-specific questions (from novice clinicians) that are necessary for differential diagnosis, our data analysis was completed in two stages. In the first stage, the reference distribution is a correlation of each cluster of questions (from the novice clinicians) making up a particular DSM PTSD Category with each (corresponding) cluster of responses from the VP representing the same DSM PTSD Category. In the second stage, variance from each individual's psychological distributions is controlled. Herein, the reference distribution reflects a semi-partial correlation controlling for the psychological factors that may be impacting the relation between each cluster of questions (from the novice clinicians) making up a particular DSM PTSD Category with each (corresponding) cluster of responses from the VP representing the same DSM PTSD Category. We also assessed the impact of absorption and immersiveness upon the "believability" of the system.

4 Results and Evaluation

4.1 Assessment of the System

Assessment of the system was completed with the data gathered from the log files in addition to the questionnaires. The log files were used to evaluate the number and types of questions that the subjects were asking, along with a measure to see if the system was responding appropriately to the questions. For a 15 minute interview the participants asked on average, 68.6 questions with the minimum being 45 and the maximum being 91. Figure 4 is a graph showing the average number of questions, asked by the subjects, lighter color, and responses by the system, darker color for each of the 8 DSM categories. It is interesting to note that most of the questions asked were either general questions (Category #G, Average 24 questions) or questions about the Trauma (Category #A, 13 questions), followed by category #C and #B, 8. The larger number of questions asked in #G was partially due to clarification questions, however we did not break down the category further to try to classify this. The distribution of questions in each category for each participant was roughly equivalent, which meant in general people asked the same kinds of questions

There are several areas in the system that can be problematic due to technological issues which would cause the system to mis-recognize the question as out of domain, something the natural language system did not know about, and generate an inappropriate response. One area was the speech recognition system. We used a speaker independent speech recognizer that did not contain all of the words or phrases asked by the subjects, as it was not known all the questions they would ask. Additionally the system did not perform as well for women voices as with men. The natural language system deals with out of domain questions by responding with an off topic response,

in our case the phrase 'I don't get what you mean'. This was a particular issue, based on the questionnaires, where the subjects got frustrated, as the system responded with this phrase too many times and there was not enough variability with out of domain responses. This response was said in total 411 times across all subjects, comparing that to the total responses of, 1066, the ratio was one in every 2.5 responses. While there is no standard for a reasonable set of questions to out of domain responses, this ratio at least gives us a measure as to how well the system was performing. While this value may seem high and did frustrate some subjects, most subjects were able to continue with questioning and get appropirate responses to perform a diagnosis. Future analysis on the speech recognition word error rate and accuracy will yield data as to what words and questions are needed to improve the speech models. It is clear from the transcriptions that the domain we built was not sufficient to capture all of the questions people were asking, the results from this study will be added to the domain for future testing. The interviewing method that people used to ask questions varied by individual; there were many different styles and personality factors that influenced the length and type of question, for example some people asked multiple segment questions, like 'hi how are you, why did you come here today?'. There are many novice assumptions by the subjects in how well this technology performs. Natural language and speech recognition is still a hard problem.

From the post questionnaires on a 7 point likert scale, the average value subjects rated the believability of the system to be 4.5. Subjects were also able to understand the patient, 5.1. People rated the system at 5.3 as frustrating to talk to, due to speech recognition problems, out of domain questions or inappropriate responses. However most of the participants left favorable comments that they thought this technology will be useful, they enjoyed the experience and trying different ways to talk to the character and also trying to get an emotional response for a difficult question. When the patient responded back appropriately to a question they found that very satisfying.

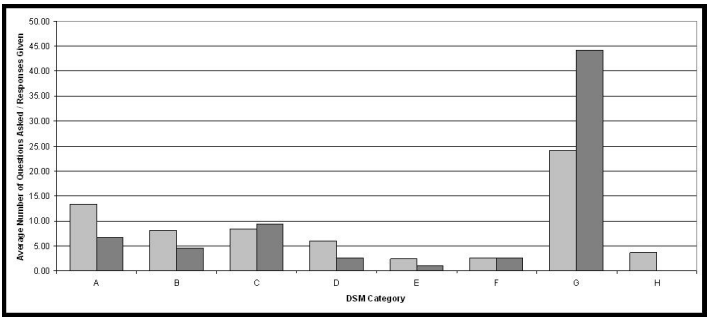

Fig. 4. Average number of questions asked, lighter, and responses given

4.2 Assessment of Student Questions and the Students

For this phase of the analysis, we aimed at investigating the relationship between a number of psychological variables and the resulting VP Responses. A summary of relations (measures as effect sizes "r") between each 1) DSM PTSD Category cluster of user questions; and 2) each (corresponding) cluster of responses from the VP

representing the same DSM PTSD Category. Please note that these are "clusters" of Question/Response pairs that reflect different diagnostic categories used for differential diagnosis.

The present focus is on effect sizes indicating strength of correlation, that is, effect sizes that describe the strength of association between question and response pairs for a given diagnostic category. Given our small sample size, we wanted a more conservative estimate of effect. Hence, an effect size (herein we use "r" as a standard of effect size) of 0.20 was regarded as a small effect, 0.50 as a moderate effect, and 0.80 as a large effect. Moderate effects existed between User Questions and VP Response pairs for Category A (r = 0.45), Category B (r = 0.55), Category C (r = 0.35), and Category G (r = 0.56), but only small effects were found for Category D (r = 0.13) and Category F (r = 0.13). After controlling for the effects of the Tellegen Absorption Scale, increased effects were found for Category A (r = 0.48), Category C (r = 0.37), Category D (r = 0.15), and Category F (r = 0.24).

We also assessed impact of psychological characteristics such as absorption and immersiveness upon the "believability" of the VP and Student interaction. To assess this relation we created a composite variable that included scores from the TAS and the ITQ (Trait Composite). Strong effects existed between the Trait Composite and the Presence Questionnaire (r = 0.78), and moderate effects existed between the Trait Composite and the Justina Post-questionnaire (r = 0.40).

5 Discussion of Results

The primary goal in this study was evaluative: can a virtual standardized patient generate responses that elicit user questions relevant for PTSD categorization? Findings suggest that the interactions between novice clinicians and the VP resulted in a compatible dialectic in terms of rapport (Category G), discussion of the traumatic event (Category A), and the experience of intrusive recollections (Category B). Further, there appears to be a pretty good amount of discussion related to the issue of avoidance (Category C). These results comport well with what one may expect from the VP (Justina) system. Much of the focus was upon developing a lexicon that, at minimum, emphasized a VP that had recently experienced a traumatic event (Category A) and was attempting to avoid (Category B) experienced that my lead to intrusive recollections (Category C). However, the interaction is not very strong when one turns to the issue of hyper-arousal (Category D) and impact on social life (Category F). While the issue of impact on social life (Category F) may simply reflect that we wanted to limit each question/response relation to only one category (hence, it may have been assigned to avoidance instead of social functioning), the lack of questions and responses related to hyper-arousal and duration of the illness (Category E) reflects a potential limitation in the system lexicon. These areas are not necessarily negatives for the system as a whole. Instead, they should be viewed as potential deficits in the systems lexicon.

A secondary goal was to investigate the impact of psychological variables upon the VP Question/Response composites and the general believability of the system. After controlling for the effects of these psychological variables, increased effects were found for discussion of the traumatic event (Category A), avoidance (Category C),

hyper-arousal (Category D), and impact on social life (Category F). Further, the impact of psychological characteristics revealed strong effects upon presence and believability. These findings are consistent with other findings suggesting that hypnotizability, as defined by the applied measures, appears moderate user reaction. Future studies should make use of physiological data correlated with measures of immersion to augment and quantify the effects of virtual human scenarios.

6 Future Work

We presented an approach that allows novice mental health clinicians to conduct an interview with a virtual character that emulates an adolescent female with trauma exposure. The work presented here builds on previous initial pilot testing of virtual patients and is a more rigorous attempt to understand how to build and use virtual humans as virtual patients along with the issues involved in building domains, the speech and language models and working with domain experts. The lessons learned here can be applied across any domain that needs to build large integrated systems for virtual humans. We believe this is a desirable application area and a small enough domain that we can perform meaningful evaluations on using VPs in real settings.

We will continue to perform more rigors subject testing with both professional medical students and with non experts to evaluate how well the different populations perform and further studies in comparing real OSCE's with real actors to the virtual patient. Investigate how incorporation of rapport [7,12] using facial analysis will further enhance the virtual patient interaction.

Additional analysis of the data includes: comparing the questions and the responses to assess how many were on and off-topic; Compute the word error rate for the speech recognition engine to assess its performance; Investigate tools to help build question/response sets for different domains; Explore ways to automate the process of classifying subject questions with the DSM categories; Build an agent framework that will be able to recognize conversation attributes such as; opening or closing statements, empathy, topic areas, follow-up and clarification questions along with more autonomous behavior, like asking questions to the clinician, assertiveness and, initiative levels, saving the conversation history and topic recognition and tracking.

People have many different interviewing and personality styles, some people are more direct, while others more empathetic. The system needs to be able to recognize these and adjust its responses. Studies are needs on incorporating learning objectives into the interview session and investigating if the virtual patient system has a learning impact is something that is valuable and will be the focus of future subject testing.

It is our belief that with adding more questions and responses the accuracy of the system will rise along with the depth of the conversions the clinician can have with the VP. In order to be effective VPs must be able to interact in a 3D virtual world, must have the ability to react to dialogues with human-like emotions, and be able to converse in a realistic manner with behaviors and facial expressions that match the clinical condition of interest. The combination of these capabilities allows them to serve as unique training and learning tools whose special knowledge and reactions can be continually fed back to trainees. Our initial goal of this study was to focus on a VP with PTSD, but a similar strategy could be applied to teaching a broad variety of

psychiatric diagnoses to trainees at every level from medical students, to psychiatry residents, to child and adolescent psychiatry residents.

Acknowledgments. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We wish to thank the Keck School of Medicine at USC, Caroly Pataki, Michele Pato, Cheryl StGeorge and Jeffrey Sugar and special thanks to Mary Slater-Kenny as the voice of Justina. This work was sponsored in part by a USC Provost grant for Teaching with Technology and the V-Humans Project.

References

1. Andrew, R., Johnsen, K., Dickerson, R., Lok, B., Cohen, M., Stevens, A., Bernard, T., Oxendine, C., Wagner, P., Lind, S.: Comparing Interpersonal Interactions with a Virtual Human to those with a Real Human. IEEE Transactions on Visualization and Computer Graphics (2006)
2. Bernard, T., Stevens, A., Wagner, P., Bernard, N., Schumacher, L., Johnsen, K., Dickerson, R., Raij, A., Lok, B., Duerson, M., Cohen, M., Lind, D.S.: A Multi-Institutional Pilot Study to Evaluate the Use of Virtual Patients to Teach Health Professions Students History-Taking and Communication Skills. In: Proceedings of the Society of Medical Simulation Meeting (2006)
3. Bickmore, T., Pfeifer, L., Paasche-Orlow, M.: Health Document Explanation by Virtual Agents. In: Intelligent Virtual Agents 2007, Paris (2007)
4. Bickmore, T., Cassell, J.: Social Dialogue with Embodied Conversational Agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. (eds.) Advances in Natural, Multimodal Dialogue Systems. Kluwer Academic, New York (2005)
5. Bickmore, T., Giorgino, T.: Health Dialog Systems for Patients and Consumers. Journal of Biomedical Informatics 39(5), 556–571 (2006)
6. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: An Architecture for Embodied C onversational Characters. In: Proceedings of the First Workshop on Embodied Conversational Characters, Tahoe City, California, October 12-15 (1998)
7. Cassell, J., Gill, A., Tepper, P.: Coordination in Conversation and Rapport. In: Proceedings of the Workshop on Embodied Natural Language, Association for Computational Linguistics, Prague, CZ, June 24-29 (2007)
8. Deladisma, A., Johnsen, K., Raij, A., Rossen, B., Kotranza, A., Kalapurakal, M., Szlam, S., Bittner, J., Sinwson, D., Lok, B., Lind, D.: Medical student satisfaction using a virtual patient system to learn history-taking and communication skills. Medicine Meets Virtual Reality (MMVR) 16 (2008)
9. DSM, American Psychiatric Association 2000 (DSM-IV-TR) Diagnostic and statistical manual of mental disorders, 4th edn, text revision. American Psychiatric Press, Inc. Washington (2000)
10. Gerhard, M., Moore, D., Hobbs, D.: Continuous presence in collaborative virtual environments: Towards the evaluation of a hybrid avatar-agent model for user representation. In: de Antonio, A., Aylett, R., Ballin, D. (eds.) Proceedings of the International Conference on Intelligent Virtual Agents, Madrid, Spain, pp. 137–153 (2001)

11. Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., Petajan, E.: Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems, 54–63 (July/August, 2002)
12. Gratch, J., Ning, W., Jillian, G., Edward, F., Robin, D.: Creating Rapport with Virtual Agents. In: 7th International Conference on Intelligent Virtual Agents, Paris, France (2007)
13. Hardin, R.M., Stevenson, M., Downie, W.W., Wilson, G.M.: Assessment of clinical competence using objective structured examination. British Medical Journal 1, 447–451 (1975)
14. Johnsen, K., Raij, A., Stevens, A., Lind, D., Lok, B.: The Validity of a Virtual Human Experience for Interpersonal Skills Education. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 1049–1058. ACM Press, New York (2007)
15. Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., Rizzo, A.A.: Virtual Patients for Clinical Therapist Skills Training. In: Pélachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 197–210. Springer, Heidelberg (2007)
16. Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D.: Building Interactive Virtual Humans for Training Environments. In: Proceedings of I/ITSEC, November 2007, Best Paper Nominee (2007)
17. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia (2006)
18. Lok, B., Rick, F., Andrew, R., Kyle, J., Robert, D., Jade, C., Stevens, A., Lind, D.S.: Applying Virtual Reality in Medical Communication Education: Current Findings and Potential Teaching and Learning Benefits of Immersive Virtual Patients. Journal of Virtual Reality (to appear, 2006)
19. Macedonio, M., Parsons, T.D., Rizzo, A.A.: Immersiveness and Physiological Arousal within Panoramic Video-based Virtual Reality. Cyberpsychology and Behavior 10, 508–516 (2007)
20. McGee, J.B., Neill, J., Goldman, L., Casey, E.: Using multimedia virtual patients to enhance the clinical curriculum for medical students. Medinfo 9(Part 2), 732–735 (1998)
21. McQuiggan, S., Rowe, J., Lester, J.: The Effects of Empathetic Virtual Characters on Presence in Narrative-Centered Learning Environments. In: Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy (to appear, 2008)
22. Norris, F.H., Kaniasty, K., Thompson, M.P.: The psychological consequences of crime: Findings from a longitudinal population-based study. In: Davis, R.C., Lurigio, A.J., Skogan, W.G. (eds.) Victims of Crime, 2nd edn., pp. 146–166. Sage Publications, Inc., Thousand Oaks (1997)
23. Parsons, T.D., Bowerly, T., Buckwalter, J.G., Rizzo, A.A.: A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods. Child Neuropsychology (2007)
24. Prendinger, H., Ishizuka, M.: Life-Like Characters – Tools, Affective Functions, and Applications. Springer, Heidelberg (2004)
25. Resick, P.A., Nishith, P.: Sexual assault. In: Davis, R.C., Lurigio, A.J., Skogan, W.G. (eds.) Victims of Crime, 2nd edn., pp. 27–52. Sage Publications, Inc., Thousand Oaks (1997)
26. Rickel, J., Johnson, W.: Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence 13(4-5), 343–382 (1999)

27. Rizzo, A.A., Pair, J., Graap, K., Treskunov, A., Parsons, T.D.: User-Centered Design Driven Development of a VR Therapy Application for Iraq War Combat-Related Post Traumatic Stress Disorder. In: Proceedings of the 2006 International Conference on Disability, Virtual Reality and Associated Technology, pp. 113–122 (2006)
28. Stevens, A., Hernandex, J., Johnsen, K., et al.: The use of virtual patients to teach medical students communication skills. The Association for Surgical Education Annual Meeting, New York, April 7–10 (2005)
29. Swartout, W., Gratch, J., Hill, R., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward Virtual Humans. AI Magazine 27(1) (2006)
30. Tartaro, A., Cassell, J.: Playing with Virtual Peers: Bootstrapping Contingent Discourse in Children with Autism. In: Proceedings of International Conference of the Learning Sciences (ICLS), Utrecht, Netherlands, June 24-28 (2008)
31. Tellegen, A., Atkinson, G.: Openness to absorbing and self-altering experiences ("absorption"), a trait related to hypnotic susceptibility. Journal of Abnormal Psychology 83, 268–277 (1974)
32. Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., Vaswani, A.: Hassan: A virtual human for tactical questioning. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007, pp. 71–74 (2007)
33. Triola, M., Feldman, H., Kalet, A.L., Zabar, S., Kachur, E.K., Gillespie, C., et al.: A randomized trial of teaching clinical skills using virtual and live standardized patients. Journal of General Internal Medicine 21(5), 424–429 (2006)
34. Walters, K., Osborn, D., Raven, P.: The development, validity and reliability of a multimodality objective structure clinical examination in psychiatry. Medical Education 39, 292–298 (2005)
35. Wessel, J., Williams, R., Finch, E., Gémus, M.: Reliability and validity of an objective structured clinical examination for physical therapy students. Journal of Allied Health 32(4), 266–269 (2003)
36. Wiederhold, B.K., Dong, P.J., Kaneda, M., Cabral, I., Lurie, Y., May, et al.: An investigation into physiological responses in virtual environments: an objective measurement of presence. In: Riva, G., Calimberti, C. (eds.) Toward cyberpsychology: Mind, cognition and society in the internet age. IOS Press, Amsterdam (2001)
37. Witmer, B., Singer, M.: Measuring presence in virtual environments: a presence questionnaire. Presence: Teleoperators and Virtual Environments 7(3), 225–240 (1998)