USING IMMERSION AND INFORMATION VISUALIZATION TO ANALYZE HUMAN-VIRTUAL HUMAN INTERACTIONS

By

ANDREW BRIAN RAIJ

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2009

For my infinitely patient wife Emily and our dog Ruby

ACKNOWLEDGMENTS

I thank my research advisor and supervisory committee chair, Dr. Benjamin Lok, for giving me the opportunity to succeed; for his advice, direction, support and patience; and for the honor of being a part of the Virtual Experiences Research Group (VERG). The people in VERG have been amazing colleagues and friends. I am proud to have worked with all of them, and I want to especially highlight Kyle Johnsen, Aaron Kotranza, Brent Rossen, John Quarles, Xiyong Wang, Robert Dickerson, Harold Rodriguez, Cyrus Harrison, Joon Hao Chuah, Lois Cao, and Dr. Yong Ho Hwang. I also want to thank Dr. D. Scott Lind, Dr. Amy Stevens, Dr. Adeline Deladisma, Dr. Diane Beck, Dr. Juan Cendan, Dr. Marc Cohen, Dr. Peggy Wagner, and all my other collaborators in medical and pharmacy education. Their perspectives have influenced my work greatly, and it has been enlightening working with them. I thank Dr. Richard Ferdig, Dr. Paul Fishwick, Dr. Chris Jermaine, and Dr. Jeffrey Ho for serving on my supervisory committee and helping me see broader perspectives on my work. I thank the University of Florida alumni who supported my work through a University of Florida Graduate Alumni Fellowship. Lastly, I thank my family: my brother Irwin; my sister Madeline; my brother-in-law Todd; the Brandon kids; my parents Jose and Elisa; my in-laws Shelly, Barbara, and Brett; my wife Emily, and our dog Ruby. I thank all of them for their love and support throughout this crazy journey to my Ph.D. They made this all possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

USING IMMERSION AND INFORMATION VISUALIZATION TO ANALYZE
HUMAN-VIRTUAL HUMAN INTERACTIONS

By

Andrew Brian Raij

May 2009

Chair: Benjamin Lok
Major: Computer Engineering

We identify verbal and nonverbal communication as the primary way in which humans try to interact with virtual human interfaces, and then use this result to develop three different approaches for analyzing interactions between humans and virtual humans. Each of these approaches is applied to analyzing interactions with virtual humans for training domain-specific interpersonal skills. In providing new approaches to analyzing virtual humans interfaces, we advance the state-of-the-art in facilitating and training interpersonal interactions with virtual human interfaces.

Verbal and nonverbal communication is identified as the primary way users try to interact with virtual humans by comparing interactions with virtual humans to similar interactions with real humans. In two user studies (n=82), participants elicited the same information from a virtual and real human using verbal communication. However, participant nonverbal behavior indicated participants were less engaged, insincere, and demonstrated a poorer attitude towards the virtual human. These behavioral differences likely stemmed from the participants' difficulty understanding the virtual humans limited expressive behavior.

The Interpersonal Scenario Visualizer (IPSViz) was then developed to enable review, analysis, and evaluation of the communication between a human and a virtual human. IPSViz generates visualizations of a human-virtual human interaction by capturing, logging, and processing the human and virtual human's verbal and nonverbal behavior.

A user study (n=27) shows that conducting an interaction with a virtual human and then reviewing that interaction with IPSViz elicits self-reflection on interpersonal skills, including verbal and nonverbal behavior, rapport-building, and communicating clearly under stress.

The next system, the Virtual Social Perspective-taking (VSP) system, enables review, analysis, and evaluation of an interaction with a virtual human from the perspective of the virtual human. The VSP system records a virtual human patients experience of talking to a medical student, and then uses the recording to transport the medical student into the patients body and relive the conversation through her eyes. The student relives the conversation to better understand the virtual human patients perspective and learn to address her and future real patients - fears. The results of a pilot study (n = 16) indicate that VSP encourages reflection on the perspectives of others and elicits self-directed change of behavior in future social interactions.

The last system, IPSVizN, enables review, analysis, and evaluation of trends and outliers in human-virtual human interactions. IPSVizN processes groups of human-virtual human interaction logs to generate summary visualizations of the interactions. An evaluation of IPSVizN with representative end-users found that participants were able to rapidly (within minutes) identify trends and outliers in overall group interpersonal skills, including verbal behavior, organization, completeness, empathy, and communicating under stress. Identifying these trends and outliers without IPSVizN would have required hours of manual effort.

CHAPTER 1
INTRODUCTION

An immersive virtual human, or embodied conversational agent [33], is a computational process that exhibits a human-like user interface. Human-like interfaces have the approximate size and appearance of a real human. They afford human-computer interaction similar to, but not necessarily the same as, human-human interaction.

The capacity of virtual humans to afford interaction similar to human-human interaction has driven the development of interpersonal simulators [86]. Interpersonal simulators simulate interpersonal scenarios (interactions with people) by replacing real humans with virtual humans. Users interact with the virtual humans to train for similar interactions with real humans. Interpersonal simulators have been developed to train a variety of domain-specific interpersonal skills, including military leadership [79], medical practice [87, 89, 115], psychology practice [155], cultural competency [10, 40], law enforcement [59], and autism education [178].

This work proposes, implements, and evaluates three different approaches to analyzing virtual human interfaces for interpersonal skills training. Each approach is built on the premise that, as in human-human interactions, the primary methods of interaction between users and virtual humans are verbal and nonverbal behavior. Thus, to analyze virtual human interfaces, one must analyze the verbal and nonverbal behavior of users and virtual humans.

Three proof-of-concept systems, IPSViz, VSP, and IPSViz^N, were developed to analyze user and virtual human behaviors. Each system demonstrates one of three analysis approaches proposed in this work.

1. **IPSViz** merges multi-perspective 3D replays of the interaction with visualizations (e.g., charts and diagrams) to give insight into speech, gaze, body posture, and other behaviors

2. **VSP** leverages immersive virtual reality to enable reliving a human-virtual human interaction as the virtual human. This allows the user to better understand the virtual human's potentially unfamiliar perspective.

3. **IPSViz^N** generates aggregate visualizations of many human-virtual human interactions to enable rapid discovery of trends and outliers in speech, body posture, tone-of-voice and other behaviors.

These approaches provide insights into user interaction with virtual humans and allow users of interpersonal simulators to get rapid feedback on domain-specific interpersonal skills that would be difficult or impossible to receive through human-human interactions. Thus, in developing and evaluating approaches for analyzing interactions with virtual humans, this dissertation provides innovations to both the study of virtual human user interfaces and their application in interpersonal simulators.

1.1 Driving Issues

1.1.1 Training Interpersonal Skills with Virtual Humans

In addition to their capacity to elicit human-like interaction, virtual humans are being applied to interpersonal skills training because they address the drawbacks of traditional role play approaches to training. The most well-known training approach is role play with real humans. In role play, trainees practice a simulated interpersonal scenario with real humans, with the goal of improving their communication in a similar, real interpersonal scenario. While role play allows for training a variety of scenarios, it has several limitations which can be addressed by augmenting role play experience with interactions with virtual humans. These limitations are:

- **Scenario diversity:** Available role play partners may not be able to simulate certain types of people, such as a person with a debilitating motor disease like Parkinson's. Virtual humans can be programmed to simulate scenarios which real humans cannot simulate easily.

- **Availability:** Role play partners are not always available to trainees whenever they need them (e.g., 3 AM before a training examination). Virtual humans can be made available to users 24 hours a day, seven days a week.

- **Standardization:** Role play partners cannot be expected to simulate scenarios in a consistent, standardized way. If a role play partner simulates a scenario 25 times, his or her behavior may unintentionally change from the first interaction to the last due to human error. Furthermore, a different role player may simulate the scenario differently. Thus, trainees who interact with the first role player could be trained

differently than those who interact with the second. Virtual humans address the standardization problem because they can be programmed to simulate a scenario the same way for all users.

- **Analysis and Feedback:** Due to the limited availability and diversity of role play scenarios, trainees cannot receive feedback often enough, nor on a diverse set of scenarios. In addition, approaches to analysis and feedback are limited to video review of or personal reflection on the interaction. The former is tedious and time-consuming, and the latter is dependent on one's memory and perspective, which may be faulty and biased. Furthermore, the analyst may need significant training to know what aspects of the interaction to analyze. This work shows how *after-action reviews* of human-virtual human experiences can be enhanced to address the analysis and feedback challenges inherent in role play experiences.

1.1.2 After-Action Reviews

In after-action reviews, trainees review an experience as soon as possible after it occurs, with the goal of facilitating reflection on one's actions and identification of strengths and weaknesses. Identifying strengths and weaknesses helps trainees emphasize strengths and improve on weaknesses in future, similar experiences. After-action reviews are routinely used in industry and the military to review real and virtual experiences [17, 18, 38, 41, 66, 72]. Their success in real and other virtual experiences motivates their use for human-virtual human experiences.

Furthermore, after-action reviews for human-virtual human interactions have the potential to provide novel feedback that would be difficult or impossible to provide through human-human interactions. Unlike human-human interactions, the computational nature of human-virtual human interactions means they can be easily logged to provide a detailed description of the interaction. The description of the interaction can then be processed to provide novel feedback to the user.

This dissertation leverages information visualization and immersive virtual reality techniques to enable rapid, focused, analysis of human-virtual human interaction logs. These analysis approaches address the challenges in analyzing real-world role play experiences, as well provide new benefits unique to interpersonal simulations:

- **Rapid feedback:** Logs can be processed rapidly to give feedback to users immediately after the interaction. Important skills and moments can be highlighted by the system, allowing users to analyze interpersonal skills more rapidly than with video review.

- **Availability:** Human-virtual human interactions and after-action review software can be made available to users at all times (e.g., with a 24 hour, 7 days a week training center). Users could download the after-action review software to review their interactions with virtual humans at any time.

- **Tightening the feedback loop:** Users can get feedback and then apply that feedback to another human-virtual human or human-human interaction minutes later. The high availability of human-virtual human experiences means this "Interaction-Review" feedback loop can be repeated ad infinitum to enable more rapid improvement in interpersonal skills.

- **Diversity:** Virtual humans can be programmed to present a wide variety of scenarios that may not be available through human-human interactions. Thus, through after-action reviews, users can get feedback on their interpersonal interaction for a more diverse set of interpersonal scenarios.

- **Overview and details:** Human-virtual human interaction can be logged at a high level of detail and then filtered and processed into overview and detail visualizations. Overviews provide an understanding of the interaction at a glance, and detail visualizations provide access to information about specific moments or interpersonal skills (e.g., gaze behavior during stressful moments).

- **Perspective:** Logs can be transformed to present human-virtual human experiences from the perspectives of others. Seeing their interaction from a new perspective can help users understand the perspectives of others, and discover aspects of their interpersonal behavior that they might not see otherwise.

1.1.3 Medical Interview Training

This dissertation explores these benefits in the context of a real-world interpersonal simulator designed to train medical professionals on interaction with patients. In the system, the user interviews a virtual human patient about his or her medical problems using natural speech and gesture, and the virtual human interacts back in a similar manner. Sensors, such as microphones, video cameras, and 3D trackers, facilitate natural interaction with the virtual human. To help the virtual human patient, the user gathers critical information from the virtual human and conducts physical examinations (through

a tangible interface connected to the simulation). Additionally, users are expected to empathize and build rapport with the virtual human patient when appropriate. After the interaction, users conduct after-action reviews of these skills (gathering information, procedure, and empathy) using the analysis approaches proposed in this dissertation.

In applying new analysis approaches to a real-world application of virtual humans, this dissertation advances the state-of-the-art in medical interview training, interpersonal skills training, and virtual human interfaces.

1.2 Thesis Statement

Reviewing interpersonal, natural interactions with a virtual human by either 1) immersing oneself in the first-person experience of the virtual human or 2) analyzing visualizations of the interactions elicits reflection on, changes perceptions of, and motivates change in domain-specific interpersonal skills.

1.3 Overview of Approach

1.3.1 Human-Virtual Human Interaction ≈ Human-Human Interaction?

Our approach to supporting the thesis statement begins with a study comparing interactions with a virtual human simulating a patient to those with a real human simulating a similar patient (Chapter 3). The goal of the study is to identify the extent to which interactions with virtual humans are similar to interactions with real humans. Establishing this similarity supports the thesis statement in two ways.

First, establishing this similarity would mean that domain-specific interpersonal skills exhibited in human-human interactions are exhibited in human-virtual human interactions. This is crucial to supporting the thesis statement because it is unlikely reviewing human-virtual human interactions would "elicit reflections on, change perceptions of, and motivate change in domain-specific interpersonal skills" if domain-specific interpersonal skills are not exhibited in human-virtual human interactions.

Second, the comparison would identify the domain-specific interpersonal skills that are important in the medical education scenarios used throughout this dissertation. Once

the skills are identified, the verbal and nonverbal behaviors associated with them can be captured and processed into a computational description of the interaction. This computational description would enable 1) generating visualizations of the interaction for review and 2) reviewing the first-person experience of the virtual human from the virtual human's perspective.

1.3.2 Capture of Verbal and Nonverbal Behavior

Once relevant verbal and nonverbal behaviors are identified, the next step in supporting the thesis statement is developing a pipeline to capture and process those behaviors for later review (Chapter 4). The pipeline is developed based on the principle that verbal and nonverbal behaviors are signals, time-varying characteristics of a human-virtual human interaction. Treating the interaction as a set of signals allows the pipeline to process and filter the interaction to compute the information needed for visualization and immersive review from the virtual human's first-person perspective.

1.3.3 Analysis of Domain-Specific Interpersonal Skills with Information Visualization Techniques

After the pipeline is developed, the next step in supporting the thesis statement is the development of proof-of-concept systems that visualize human-virtual human interactions, and determining if these systems elicit reflection on, change perceptions of, and motivate change in domain-specific interpersonal skills. Two after-action review systems are developed, the Interpersonal Scenario Visualizer (IPSViz), and IPSViz^N.

1.3.3.1 IPSViz

The first system, IPSViz (Chapter 4), generates overview and detail visualizations of the verbal and nonverbal behavior in a human-virtual human interaction. With a glance, the overview visualizations provide a rapid sense of how domain-specific interpersonal skills are used in an interaction. Specific medical domain skills that are presented with overview visualizations include user gaze behavior, the organization and thoroughness of the user's questions, and the user's interpersonal distance to the virtual human.

Interactive tools allow users to drill-down and review details about the interaction, including identifying specific moments to review from a detailed list of user and virtual human speech and video playback of those moments.

To evaluate IPSViz, health professionals interact with a virtual human patient and then conduct an after-action review (AAR) of their interaction using IPSViz. Pre- and post-AAR surveys are administered to participants to measure the change in their perceptions of their interpersonal skills due to using IPSViz. In addition, participants are interviewed to collect subjective feedback about the AAR. If there is a significant change in perceptions of one's domain-specific interpersonal skills, and/or participants discuss lessons learned from the review session, then it is clear participants self-reflected on, changed perceptions of, and considered changes in domain-specific interpersonal skills.

1.3.3.2 IPSViz^N

IPSViz^N is similar to IPSViz in that it also presents visualizations of human-virtual human interactions (Chapter 6). However, IPSViz^N aggregates interaction signals from *many* human-virtual human interactions, rather than just one, to present overviews and details of *group* interpersonal skills. These visualizations enable users to rapidly identify trends and outliers in the domain-specific interpersonal skills of groups.

A pilot study evaluates the benefits of IPSViz^N for a group of health professions educators. The health professions educators are representative of a class of users who do not interact directly with the virtual human, but instead direct students to conduct human-virtual human interactions. These users need to reflect on the human-virtual human interactions they design and direct, identify trends and outliers in the communication skills of their students (perceptions), and make changes in their student's future interactions with virtual and real humans. In the study, multiple versions of IPSViz^N are distributed to participants, with feedback collected from participants after each version. Each version incorporates the previous version's feedback to improve the

tool's ability to elicit reflection on, change perceptions of, and motivate change in the domain-specific interpersonal skills of groups.

1.3.4 Analysis of Domain-Specific Interpersonal Skills with Immersive Virtual Reality Techniques

The last step in supporting the thesis statement is the development and evaluation of the Virtual Social Perspective-Taking (VSP) system, a proof-of-concept system for immersing oneself in the first-person experience of a virtual human (Chapter 5). The VSP system allows a user wearing a head-mounted display to see what a virtual human saw, as well as hear what the virtual human heard by playing back audio of from the logs. In addition, the VSP system prompts the user to act out the virtual human's behaviors (speech and movement). This immersion in the virtual human's first-person experience is aimed at helping the user better understand the virtual human's perspective.

The VSP system is evaluated by measuring how virtual social perspective-taking changes a user's perceptions of 1) a virtual human patient interaction the user conducted and 2) a future virtual human patient interaction the user conducts after the VSP interaction. First, the user interacts with a virtual human patient, Then, the user reviews that interaction using the VSP system. Finally, the user conducts another virtual human patient interaction. A survey is given after each patient interaction and the VSP interaction. This study design allows determining if immersion in a virtual human's first-person perspective elicits self-reflection on, changes self-perceptions of, and motivates change in domain-specific interpersonal skills.

1.4 Innovations

The innovations this dissertation provides has broad impact on both the science of simulating interpersonal scenarios with virtual humans and their emerging application for training interpersonal skills. The innovations provided by this dissertation are:

- The further establishment of the similarity between interpersonal simulations and the real-world experiences these simulations are based on.

- The identification of verbal and nonverbal communication, the primary method of interaction in human-human interactions, as the primary method of interaction afforded by human-virtual human interactions.
- A pipeline for computationally capturing and processing verbal and nonverbal communication for after-action review.
- The pairing of after-action review and human-virtual human experiences to gain insights into the communication skills of oneself and others. Some of these insights would be difficult or impossible to gain from after-action reviews of human-human interactions.
- Systems for reviewing human-virtual human interactions using information visualization techniques.
- A system for reliving a human-virtual human interaction from the first-person perspective of the virtual human.
- Through after-action review, human-virtual human experiences can elicit reflection on interactions with virtual humans, and self-directed change in future interactions with real humans.

1.4.1 Interpersonal Simulations ≈ Real-world Interpersonal Interactions

This dissertation innovates by establishing the similarity between interpersonal simulations and the real-world experiences these simulations are based on. Unlike previous work establishing this similarity, this dissertation compares an interpersonal simulation to *a real-world counterpart*. Using a real-world counterpart provides a gold-standard to which to compare the interpersonal simulation and establish its similarities and differences.

The real-world counterpart was a medical interview, an interpersonal interaction where doctors must communicate effectively with patients to diagnose and treat them. Interviews between medical students and an actor playing a patient were compared to interviews between medical students and a similar virtual human patient. The comparison indicated that the simulated interpersonal interactions approximated real-world interpersonal interactions. This approximate equivalence bolsters the argument that interpersonal simulations can be used for teaching, training, and evaluating real-world interpersonal skills.

1.4.2 Verbal and Nonverbal Communication

The comparison between the interpersonal simulation and a real-world interpersonal interaction also highlighted that, much like with real humans, verbal and nonverbal behavior are the primary mechanisms by which users communicate with virtual humans. Study participants used verbal behavior with the VH that was similar to that with real humans. Participants talked to the virtual human and the virtual human talked back to the participant. Participants also verbally expressed empathy to the virtual human, and tried to build rapport with the virtual human.

While participants also used nonverbal behavior with the VH, that behavior communicated different messages than nonverbal behavior with the real human. With the virtual human, users nodded less, had a more distant posture, and used a more monotone voice. These nonverbal signals communicated a poor attitude, low interest, and insincerity towards the virtual human.

The extensive use of verbal and nonverbal communication with virtual humans, and its similarity to that with real humans, implies that analyzing human-virtual human communication provides insight into human-human communication. This further motivates the computational analysis of human-virtual human communication. While standard, video-based analysis of human-human communication is possible, sensor-driven, natural virtual human interfaces afford a more sophisticated approach.

1.4.3 Capture, Process, and Display

Leveraging human-virtual human interaction sensors, an innovative pipeline was developed to capture, process, and display human-virtual human communication. The pipeline has three stages capture, process, and display. In the capture stage, human communication with a virtual human is captured with sensors (e.g., head trackers and speech recognition software) and logged. The virtual human's communication with the human, which is generated by the interpersonal simulator, is also logged. The logged interaction data is treated as an *interaction signal* which must be filtered and processed

before it can be visualized. Filtering and processing 1) compensates for sensor error exhibited in the interaction signals, 2) fuses interaction signals to form new signals, and 3) extracts new signals which are embedded in other signals. After filtering and processing, interaction signals are ready for display. Display can take the form of abstract visualizations (such as plots of interpersonal distance, and transcripts of the interaction) or literal representations of the interaction (such as video and 3D reconstructions). This enables review and analysis of the spatial, temporal, and social characteristics of verbal and nonverbal communication with a virtual human. As verbal and nonverbal communication with a virtual human is reflective of verbal and nonverbal communication with a real human, this pipeline enables gaining insight into human-human interpersonal skills using human-virtual experiences. This dissertation leverages this pipeline to enable a specific method for gaining insight into interpersonal skills, the after-action review.

1.4.4 Insight through After-Action Review

Pairing after-action review with human-virtual human experiences facilitated gaining insights into the real-world interpersonal skills of oneself and others. Three after-action review systems were developed to explore the impact of after-action review for human-virtual experiences, IPSViz, VSP, and IPSViz^N. These systems leverage the previously described pipeline, information visualization, and immersive virtual reality techniques to allow users to review human-virtual human experiences in novel ways. Collectively, these systems allow users to gain insight into human-human interpersonal skills from human-virtual human interaction.

IPSViz and VSP allowed users who interacted with a virtual human to gain insight into their own communication skills. With IPSViz, users self-reflected on their communication with a virtual human and identified ways they might change behavior with real humans. Specific changes users proposed include improving gaze behavior, conducting a more focused verbal interaction, and being more empathetic to the virtual human. After using the VSP system, users indicated they better understand the virtual human's

perspective. They felt they did not communicate confidence, were not empathetic enough, and did not address the virtual human's concerns and fears well.

IPSViz^N allowed users who monitor human-virtual human interactions to identify trends and outliers in the communication skills of groups. Trends discovered among a group of medical professionals who interacted with a virtual human patient include more flexibility among more skilled medical professional than less skilled counterparts and a tendency for skilled physicians to discuss family history less. Outliers in the same group represented medical professionals who never showed empathy, were too brief, and/or did not discuss important topics with the virtual human. From these trends and outliers, after-action review participants suggested remedial training for the outliers and considered how to use the tool to deepen their understanding of what communication skills lead to successful interaction with patients.

1.4.5 After-Action Review + Virtual Human Experiences = Self-directed Change

By proposing after-action reviews for virtual human experiences, this dissertation expands the capabilities of virtual human experiences. Virtual human experiences are not only capable of providing practice of interpersonal interaction [87], but can also encourage self-reflection on these interactions. This self-reflection can lead to self-directed change in interpersonal interactions with real humans. By interacting with a virtual human, and conducting an after-action review of that experience, interpersonal interactions with real humans can be improved.

CHAPTER 2
PREVIOUS WORK

This dissertation is influenced by, and provides innovations back to, four areas of research: virtual humans (Section 2.1), after-action reviews (Section 2.3), temporal visualizations (Section 2.4), and visualization of interpersonal communication (Section 2.5).

2.1 Virtual Humans

This section reviews previous research on virtual humans, placing them in the context of this dissertation's innovations. Four questions about virtual human experiences are addressed: *What is a virtual human? What are the characteristics of effective virtual humans? In what contexts have virtual humans been used? Are virtual humans similar to real humans?* Of particular relevance for this dissertation is their effectiveness as surrogates for real humans, and their emerging use as conversation partners.

2.1.1 What Is a Virtual Human?

For the purposes of this dissertation, virtual humans, or embodied conversational agents [33], are computational processes that exhibit a human-like interface. A human-like interface means that the virtual human approximates the appearance, perception, and behavior of a real human. The extent to which virtual humans must have these characteristics, as well as the best ways to simulate these characteristics, remain open research problems [189]. This dissertation focuses on a different problem: building effective virtual human experiences for interpersonal skills training.

2.1.2 Effective Virtual Humans

Researchers have worked to establish the basis of effective virtual humans. Badler et al. [13] suggest virtual humans "should move or respond like a human" and "must exist, work, act and react within a 3D virtual environment." Alessi and Huang [6] expand these rules for psychology applications. They suggest virtual humans should be social, emotionally expressive, and interactive. Virtual humans should "capture and assess a

viewer and their emotional status, then translate this, taking into consideration cultural, educational, psychosocial, cognitive, emotional, and developmental aspects, and give an appropriate response that would potentially include speech, facial, and body emotional expression."

Thórisson and Cassell [181] agree that human-like emotional expression is important, but non-verbal behaviors that support conversation, e.g., gesturing at objects and looking at the user to indicate attention, are more significant. Cassell et al. [33] focus on modeling conversational interaction to create believable, functional virtual humans. Blascovich [25] theorizes virtual humans with higher behavioral realism are more persuasive. Vinayagamoorthy et al. [190] suggest a virtual human's expressions and behavior should be appropriate for the application's context.

There is also significant evidence that making virtual humans more human-like does not necessarily make them more effective or acceptable to users. This notion was first proposed for robots by Mori [126] as the Uncanny Valley theory. The Uncanny Valley theory argues that when a virtual human reaches a certain threshold of human-realism, reactions to the virtual human become negative. Vinayagamoorthy et al. [188] found study participants exposed to different types of virtual humans were more forgiving of flaws in less realistic, cartoon-like characters. Steptoe et al. [172] found a similar tradeoff between improving the fidelity of a virtual human's eye gaze behavior and the ability of users to accurately judge a virtual human's communication.

Other work indicates virtual humans should have human-like characteristics with *consistent* levels of fidelity. Garau et al. [65] found that the quality of communication between a virtual human and real human was rated higher when the virtual human's appearance and behavior were at similar levels of realism. Similarly, Bailenson et al. [16] found that copresence (the sense of being with another person in a virtual environment) with a virtual human was lowest when a virtual human's behavioral and

visual fidelity were inconsistent. An extensive review of the virtual humans literature led Vinayagamoorthy et al. [190] to similar conclusions.

2.1.3 Virtual Human Applications

Despite these open problems, virtual humans have been effectively applied in a variety of scenarios. The University of Southern California's Institute for Creative Technologies created virtual human experiences to train soldiers in interpersonal leadership [79]. The Just VR system [115] allows a medical trainee to work with a virtual assistant to treat a virtual victim. Balcisoy et al. [19] created a system where users play chess against a virtual human augmented to the real world. The Virtual Classroom [155] uses virtual teachers and students to assess attention and social anxiety disorders. Babu et al. [10] use virtual humans to teach South Indian greeting etiquette, and Deaton et al. [40] uses virtual humans to teach cultural competency to soldiers working in foreign lands. Frank et al. [59] use virtual humans to teach law enforcement personnel how to interact with the mentally unstable. The Virtual Experiences Research Group at the University of Florida uses virtual humans to train medical students on communicating with patients [87, 89]. Ryokai et al. [159] uses virtual humans as storytelling partners for young children. Thórisson's [180] interactive guide, Gandalf, gives tours of the solar system. The Human Modeling and Simulation Group at the University of Pennsylvania uses virtual humans for task analysis and assembly validation [12]. Tartaro et al. [178] are using interactions with virtual humans to help people with autism spectrum disorders develop interpersonal skills. The common theme in almost all these applications is the use of virtual humans as surrogates for real humans.

2.1.4 Virtual Humans as Surrogates for Real Humans

Several studies provide the theoretical basis for using virtual humans as surrogates for real humans. These studies highlight that users interact with and are affected by virtual humans as if they are real humans.

Bailenson et al. [14, 15] have shown that people manage personal space with virtual humans in the same manner as they do with real humans. Study participants kept more distance from a high-status virtual human than from an unknown virtual human and more distance from a virtual human than inanimate virtual objects. Female participants kept more distance from virtual humans that maintained eye contact than with virtual humans that did not.

Zanbaka et al. [201] have shown virtual entities (human or animal-like) can be as effective as real people at persuasion. In interactions with both real and virtual speakers, persuasion was stronger when participants listened to a speaker of the opposite-sex. In a different study, Zanbaka et el [202] found that, as with real humans, the presence of a virtual human lowers performance on novel or complex tasks.

Pertaub et al. [139] noted participants with a fear of public speaking reported similar anxieties when speaking to an audience of virtual people. Garau et al. [64] showed that people represented by avatars communicate better when the avatars employ realistic, task-appropriate eye-gaze.

These characteristics have been observed for other forms of computational agents. For example, Nass et al. [130] explored the affective power of computers and intelligent agents. Their work has shown people ascribe human characteristics to computers, such as helpfulness, usability, and friendliness.

2.2 Metacognition: Monitoring, Reflection, Evaluation, and Change

Using interactions with virtual humans, this dissertation facilitates *metacognition* for domain-specific interpersonal skills. Metacognition [9, 27, 55, 154, 203] is the act of self-regulating learning by:

1. **Self-monitoring** one's actions, thoughts, and learning process
2. **Self-reflecting** on and **self-evaluating** those actions, thoughts, and learning processes

3. **Self-directing change** of actions, thoughts and learning processes in future situations

Good metacognitive skills are hypothesized to be an indicator of academic, professional, and personal success. Metacognition is associated with positive educational outcomes, self-confidence, and happiness [153]. Crucial to the metacognitive process is self-awareness. A person must understand herself to self-monitor, self-reflect, and self-evaluate accurately and with confidence. Increasing self-awareness can significantly improve metacognition [39, 48, 184]. Thus, tools used to facilitate metacognition place a heavy emphasis on self-awareness.

The most common and most researched tool used for metacognition is video. Reviewing video facilitates metacognition by allowing users to see interactions from the perspective of an external observer. Video review has had positive impacts in many domains. For example, video review can improve self-perceptions of performance on a speech therapy exercise [122], decrease aggression in adolescents [133], improve social skills of children with autism [179], medical interview skills of medical students [136, 157], and self-understanding for people who are socially anxious [149]. In all of these studies, the video review made people reevaluate their self-perceptions and change actions and strategies in future situations.

Reflective writing has also been used extensively for metacognition. The Learning Assessment Journal [31] is a set of forms, each geared towards a different pedagogical goal, which guides and structures student writing. In a study, the structured writing forced students to reflect on themselves and their learning process. Action logs [129] are daily diaries students write to self-reflect on the day's lessons. Teachers can also read action logs to facilitate their own self-reflection on how to improve teaching and lessons. McCrindle et al. [120] tested self-reflective writing in a first-year university biology class. Participants were either tasked with writing a single report at the end of the semester (the control group) or writing journal entries periodically throughout the course. Those who wrote

in the journal received significantly better grades than those in the control group, and demonstrated a deeper awareness of their learning process and problem-solving strategies.

Lastly, graphing has been used to facilitate metacognition. Graphs are typically used to present objective performance data and milestones. To improve self-awareness, people compare the objective graphs to their own self-biased perceptions of their performance. They also use graphs to track performance over time and see if they meet milestones. Reviewing graphs has helped students with learning disabilities perform cognitive tasks better and be more productive [45], improved motor skills, self-esteem, and self-awareness of learning [95], and helped teachers become more self-aware of their teaching mistakes [92].

2.3 After Action Review

This dissertation's primary innovation is the extension of virtual human experiences with after action reviews (AARs). AARs are structured reviews of an event (e.g., a military training exercise or a business meeting). AARs facilitate metacognition by helping event participants reflect on their choices and consider alternatives for similar, future events (e.g., a military battle or the next business meeting).

This section places this dissertation in the context of existing knowledge about AARs. Both the theory and practice of AARs are discussed, with a focus on AARs for virtual experiences. Although virtual experiences have incorporated AARs previously, to our knowledge, this work represents the first use of AARs for virtual human experiences. Virtual human experiences are characterized by significant interpersonal communication, and thus require different approaches to AAR than other virtual and real experiences. This dissertation identifies several approaches to AARs of VH experiences, expanding what is known about AARs and identifying their usefulness for virtual human experiences.

2.3.1 Fundamentals of After Action Review

The after action review (AAR) was originally developed by the United States military to improve the training and evaluation of soldiers [17]. Successful AARs encourage

self-reflection, self-discovery, and self-directed change. They allow "soldiers to discover for themselves what happened, why it happened, and how to sustain strengths and improve on weaknesses" [41]. AARs have been so effective in the military that other organizations (e.g., businesses, not-for-profits, and government agencies) incorporate AARs into their employee training and evaluation programs [18, 38, 66, 72].

One aim of this dissertation is to demonstrate *effective* after action reviews for virtual human experiences. Hence, this section explores the characteristics that make AARs effective. Furthermore, this section also identifies corresponding challenges that must be overcome to extend virtual human experiences with after-action reviews. The characteristics and challenges identified below are derived from a meta-analysis of the following resources on after action review [17, 18, 38, 41, 66, 72].

- **Timing:** AARs should be conducted as close to the events as possible. In some situations, it is appropriate to conduct AARs during the event itself. This insures memories of the event are fresh in the minds of the participants.

- **Desired outcomes:** The AAR begins with identifying what the desired outcomes of the event were. Identifying desired outcomes provides a starting point for discussion. AAR participants can then determine if these outcomes were met by discussing what actually happened.

- **Objective description:** Event participants should describe what happened during the event in their own words. This helps participants understand how others view what happened. Through this discussion, a consensus description of the event is established. The consensus description should be objective. It should describe the facts of what happened. This insures that lessons learned from the AAR are not based on personal biases or opinion, but on the reality of what happened.

- **Training aids:** Training aids, such as 3d models (physical or digital), maps, charts, and video, should be used during the description phase. Training aids help participants describe what happened in context of the spatial and temporal features of the event. Spatial descriptions identify the event's setting and the location of participants and objects individually and with respect to each other. Temporal descriptions highlight key moments to review and establish the order of actions during the event.

- **Identifying causality:** A major goal of the AAR is to identify causality, how participant actions and event constraints (e.g., the location of the event) led to specific outcomes (that may or may not align with the desired outcomes of the

event). Once causality is identified, participants can decide how actions should be changed in future events to meet desired outcomes.

- **Targeted review:** AARs should focus on specific, targeted areas to improve, rather than a broad review of an entire event. This helps participants identify specific changes in behavior for future scenarios.

- **Disseminating lessons learned:** The lessons learned from after-action review should be disseminated throughout an organization. By disseminating lessons learned, those who did not participate in the event can also learn from the event, and apply those lessons in new contexts. Disseminating knowledge allows small-group AARs to create large-scale, beneficial change throughout an organization.

- **Success:** Ultimately, the success of an AAR is measured on whether it 1) changes future actions of participants and organizations in 2) ways that lead to desired outcomes. If the desired outcomes are reached in future events, then the lessons of the AAR are validated.

These guidelines for successful AARs highlight several challenges for developing AARs for virtual human experiences.

- How soon after a virtual human experience can an AAR occur?

- How can the participant gain an understanding of how the virtual human understood the experience?

- How can an AAR system encourage objective description of the VH experience?

- What are appropriate training aids for AARs of virtual human experiences? Can training aids be developed to help AAR participants identify causality?

- Can an AAR of a virtual human experience change future actions of participants and lead to desired outcomes?

- How can lessons learned from an AAR be disseminated to those who did not participate in the AAR?

These challenges are addressed in this dissertation through the design and evaluation of three after action review systems for virtual human experiences, IPSViz, IPSViz^N, and VSP.

2.3.2 After Action Review Systems

Existing AAR systems facilitate review of a diverse set of scenarios, including medical training, military training, emergency response training, virtual experiences, user interaction, and scientific experiments.

2.3.2.1 Applications of AAR systems

Militaries and emergency response forces are the most common users of AAR systems. Military AAR systems enable review of training exercises, including operations in urban terrain [83], direct-fire, troop-to-troop combat exercises [112], large-scale battle and war simulation [109, 119, 137, 169], flight simulations [54, 71, 152], team communication [102], and dismounted infantry simulations in virtual environments [97]. In addition to training exercises, militaries have used AAR systems to review real military engagements that have already occurred [125]. Emergency response AAR systems have been built for reviewing chemical leak and spill cleanup and rescue operations [85, 128], as well as for communication and coordination among fire fighters [67]. These systems support review of live, virtual, and mixed reality training exercises.

AAR systems are also seeing increased use in the medical domain. Laparoscopic and other surgical simulators often include tools for reviewing training exercises [63]. The Roter Interaction Analysis System (RIAS) [157] allows medical students and practitioners to review and rate patient-doctor interactions. Quarles et al. [142] added AAR functionality to a mixed reality training system for anesthesia machines. This dissertation presents an AAR system for patient-doctor interactions, where the patient is a virtual human.

Finally, AAR systems have been used to review user experiences in virtual experimental virtual environments and games. Phloem [127] allows psychology and virtual environments researchers to review scientific experiments conducted in virtual environments. Steed et al. [171] built an AAR system to study breaks in presence. (A break in presence is a moment when a user in a virtual environment becomes aware of being in a real environment.)

Kim et al. [94] instrument video games to collect user interaction data, and then conduct visualization-assisted AARs of the data to learn how to improve games.

In addition, review systems have been built into core virtual reality (VR) tools and platforms, such as tracking systems, toolkits, and game engines [58, 105, 121, 131]. These core VR tools provide basic AAR functionality to the applications built on these tools, as well as better tools for diagnostics (i.e. debugging).

2.3.2.2 Common characteristics of AAR systems

While the scenarios for which these AAR systems were built are diverse, these AAR systems share three common characteristics: *capture*, *display*, and *evaluation*.

Capture: Capture is the recording of an event using sensors (e.g., video or motion tracking sensors). The sensor data represents an objective description of the event. This description can be processed and analyzed (computationally or by hand) to gain insight into the event.

Many types of sensors are used for capture. The most common sensor is the video camera, which provides a video recording of an event [83, 94, 157]. 3D sensors are used to track the positions and orientations of vehicles, people, and tools (e.g., tanks, soldiers, and weapons) [83, 97, 170]. Physiological sensors (e.g., heart rate and galvanic skin response monitors) capture information about user stress levels or emotions [83]. Microphones, speech recognition software, e-mail logs, and text-message loggers capture communication between event participants [102].

For large-scale distributed events, AAR data is captured by a heterogeneous, distributed, sensor network. Significant infrastructure is needed to aggregate, distribute, and synchronize the heterogeneous, distributed data for visualization, playback, and evaluation [56, 57, 163].

Several approaches to recording a virtual experiences (VE) for later review have been proposed. In Hart's image-based approach [74], a cube is aligned with the VE user's viewpoint. Each face of the cube represents the near plane of a camera frustum

facing away from the cube. Each camera's view is recorded to form an environment map. To review the VE, the environment map is reprojected onto a viewer-centered cube. Greenhalgh et al. [69] propose *temporal links* that define a relationship between present and past events in a VE. Defining the temporal, spatial and presentational aspects of this relationship enables adding past events in a VE into a VE occurring in the present. Friedman et al. [60] propose standardizing the recording of VEs for analyzing presence. Guidelines are derived by analyzing a real presence data set. An important distinction is made between temporal data (e.g., system events, tracking, video recordings) and nontemporal data (e.g., questionnaires). The authors propose building tools that aggregate VE data to replay or summarize VEs.

AAR systems presented in this dissertation aggregate temporal data associated with a VH experience (e.g., tracking data, audio, video, speech recognition data, event logs, and VH behavior) to produce visualizations of the interaction. The visualizations characterize the communication between a VH and a person. Characterizing this communication enables AAR of VH experiences for communication skills training.

Display: Display is the transformation of captured sensor data into visuals, such as charts, graphs, maps and animations. Maps, building layouts, 3D renderings and other spatial visualizations enable displaying events in the space that they occurred [56, 57, 97, 109, 119, 127, 137, 142, 163, 169–171]. This provides context and helps demonstrate how spatial characteristics led to specific outcomes. Timelines (events and training variables graphed versus time) show how events unfold and the causal relationships between events (e.g., how did an event at time t cause a later event at time $t + n$) [102, 109, 119, 127, 137, 157, 169]. Playback is a form of visualizations that displays one sample at a time, but in rapid succession (animation). Common forms of playback are video, 2D animation, and 3D animation. By watching a playback of an event, AAR participants learn how event outcomes unfold over time. As playing back an entire interaction for review can be time-consuming, AARs often index recorded events

and display them in a timeline or transcript. Users choose moments from the timeline or transcript to play back those moments, thus enabling more rapid review. Teachable moments [173] - everyday life moments where there is an opportunity to teach a lesson - can be highlighted on transcripts and timelines, allowing users to review those moments which are most important.

Much of the data associated with AARs are inherently temporal, in that they describe how the captured event evolves over time. The visualization of temporal information is an important sub-class of visualization that is discussed in more detail in Section 2.4.

Evaluation: Evaluation is the analysis of the captured data describing the event. The aim of evaluation is to guide users towards the lessons to be learned from the AAR. Lessons can be determined through manual, human analysis and annotated to the playback or visualizations [157]. In addition, evaluation is increasingly being done computationally. Computational approaches to evaluation aim to automatically identify important moments to review, determine causality, and provide targeted, pedagogically-focused feedback to participants.

The Automated Communications Analysis System (AutoCAS) [102] processes verbal human-human team communication to provide needed communication skills feedback [96]. Speech recognition software transcribes communication between teammates, enabling display of communication transcripts, categorization of speech utterances, and a timeline that shows when different participants are speaking. Although the speech recognition process produces noisy transcriptions, the system helps participants understand how poor team communication can lead to undesired outcomes.

Mao and Gratch [68, 117, 118] have developed a system to determine causality from logs of an event, and then provide this data back to event participants in an AAR. Given an outcome, the system back traces through event logs to the individuals or conditions that led to an outcome. Identifying causality enables AAR systems to automatically identify "teachable moments" for review.

The Adaptive Instructional System Intelligent Flight Trainer (AIS-IFT) [152] is an intelligent tutoring system for a helicopter simulator. The tutor analyzes user behavior and provides feedback both during and after the simulation. After the simulation, AIS-IFT summarizes user task performance and provides replay functionality. The AAR is tailored both to the simulation's pedagogical goals and to the user's history with the simulator. For example, students do not review skills during the AAR which their history indicates they've already mastered.

The DARWARS project [114] automatically analyzes simulation logs to determine if participants met the pedagogical goals of the simulation. Pedagogy is defined with three components: mission objectives, constraints, and measures of success. These are defined at multiple scales, enabling AARs for individuals, small teams and large teams. Tracking measures over time allows AARs to show trainee progress and predict trainee performance in the real world.

2.4 Visualization of Temporal Data

This dissertation is influenced significantly by the field of information visualization and its sub-field, temporal visualization. Temporal visualization is the transformation of time-series and other temporal data into visual displays [2, 3, 7, 37, 164]. As with most visualizations, temporal visualizations aim to help users answer questions or discover unknown properties of the underlying dataset that would be time-consuming, difficult or impossible to determine otherwise. Visualization has been used to analyze a variety of temporal datasets including health [70, 140] and juvenile detention records [140], web site [82] and application [76] logs, server security logs [177], photo histories [84], migration paths [174], hotel visitation records [193], film box office records [29], and music listening logs [29]. Temporal visualizations are used in this dissertation to help users gain insight into human-virtual human interactions.

The rest of this section discusses temporal visualizations from several perspectives. First, four types of temporal visualiztions are described: Instant, Interval, Period, and

Branching Time visualizations [2, 2, 164]. Then, a special class of visualization that combines temporal data with spatial data is discussed. Lastly, the roles of interaction and coordination in visualization are described.

2.4.1 Instant and Interval Visualizations

Instant visualizations depict a specific moment t and interval visualizations describe a specific time range $[t_1, t_2]$ in a stream of temporal data. The most common type of instant and interval visualization is the timeline. Timelines plot intervals and instants on an axis representing the temporal dimension (typically the horizontal) and their corresponding values are plotted on a second axis (typically the vertical. Without loss of generality, we assume this throughout this section) [70, 76, 98, 140, 182]. Intervals are often represented graphically with horizontal bars whose left and right boundaries correspond to the edges of the time interval t_1 and t_2. Instants are often depicted with points or glyphs, icons or pictorial images that depict the meaning of the data underlying the corresponding instant [140]. Multiple time series can be plotted on the same timeline to facilitate comparison and highlight relationships between datasets.

Stacked plots [29, 75] are a special type of timeline visualization that use area to depict the value of a temporal variable. Stacked plots are typically used to depict relative quantities or percentages. Variables are displayed on the plot as 2D regions. The regions are stacked one on top of the other (assuming time is the horizontal axis), and the size of each region shrinks and grows vertically depending on the value of the underlying variable at different instants of time. This allows for rapid comparison of the relative "strength" or contribution of a set of variables.

Another form of instant visualization is the snapshot. Snapshots display the state of a variable at a specific moment in time t using a single image. The difference between a snapshot and a timeline is that a timeline will usually display multiple instants for a variable, whereas a snapshot only depicts the single instant alone [49]. Often, snapshots

are strung together and animated to help users see how a set of variables evolve over time [1].

2.4.2 Periodic Visualizations

Periodic visualizations display temporal data using fixed intervals or cycles. The most well-known periodic visualization is the Gregorian calendar [21, 113, 185, 193]. A Gregorian calendar is a table where each row represents a week, each column represents the repetitions of each day of the week, and each cell represents a specific day. Variations on the Gregorian calendar (sometimes called reruns [193]) change the period of the row or column (e.g. from 7 to 14 days), or change the level of detail by, for example, zooming in to a week and depicting cells as hours, or zooming out to a year and depicting each month as a cell. When temporal data is plotted on a calendar, the vertical and horizontal alignment of periods highlight cyclical patterns in the data.

Another common cyclical visualization is the spiral [22, 30, 194]. Spirals can be thought of as a timeline bent into a series of concentric circles, where each circle represents a period of a cycle. Just as with timelines, dots, glyphs, and bars (instants and intervals) can be plot on the circles. The radial alignment of the circles highlights cyclical patterns in the data.

2.4.3 Branching Time Visualizations

Branching time visualizations depict a temporal data stream which can diverge into multiple temporal data streams (i.e. alternative timelines). The most common uses of branching time visualizations are depicting uncertain future events and the predicted consequences of a set of decisions [3]. PlanningLines [4] depict uncertainties in intervals using horizontal bars marked with uncertain start and end times. VizTree [106] depicts alternative timelines as a tree structure.

2.4.4 Spatio-Temporal Visualizations

A class of visualizations related to temporal visualizations are spatio-temporal visualizations [7]. Spatio-temporal visualizations display data with both spatial and

temporal dimensions simultaneously. The most common spatio-temporal visualizations are geographic in nature. They incorporate maps with temporal data. For example, temporal paths can be drawn along the maps (e.g., migration patterns in Europe [174]), or animation can be used to depict how characteristics change across a geographic region over time [111]. Other spatio-temporal visualizations depict space as two dimensions of a cube and time as the 3rd dimension. Space-time cubes [110] are analogous to taking the frames in a video and stacking them on top of each other to form a cube.

2.4.5 Interaction with Visualizations

"Overview first, zoom and filter, then details on demand" [161] is a philosophy proposed by Shneiderman describing how visualization tools should be structured and what kinds of interaction facilities such tools should include. This philosophy extends to temporal visualization tools. First, users should be presented with a visual overview of a temporal dataset. A good overview allows users to get the gist of the data, or overall trends in the data, with a glance. In addition, a good overview also helps users identify features in the data to explore more closely. Features can be explored in detail by adding panning and zooming facilities to the interface [3, 98, 182]. Filtering is also useful for removing uninteresting data from the display. For example, Timeboxes [80] allow users to select a series of overlapping boxes, and information that only meets the parameters of the box are displayed. Similar in principle to filtering, queries can be specified to search for or extract temporal patterns of interest from a dataset. Timeboxes can be drawn over a temporal pattern to find other similar patterns in a dataset [28, 80]. Another query approach provides users with tools to draw a temporal pattern to search for in the dataset [53, 192]. After zooming, panning, filtering, and querying, users should be able to interactively display details on demand, specific information about individual data points.

One challenge with the "Overview first, zoom and filter, then details on demand" philosophy is that users can lose track of context as they drill down into the details. To resolve this issue, visualization tools incorporate "focus+context" visualizations

[62]. Focus visualizations show the details that the user is interested in. On the same screen, context visualizations display the overview and highlight the range of the focus visualization on the overview. One common approach to "focus+context" is the application of spatial distortions to the visualization, such as a logarithmic scale on a line plot or a fish-eye lens distortion on a very large graph structure [61].

2.4.6 Coordinated Visualization

Sophisticated temporal visualization tools incorporate multiple visualization types and interaction approaches [76, 162]. These tools link the visualizations together such that interaction with one visualization affects the others. For example, users might select a specific instant in a timeline, and then a graphical depiction of that instant in another window is highlighted. Similarly, users might select a set of interesting points on a line plot, and related spatial data might be highlighted on a map in another window. This linking of multiple visualizations through interaction is known as coordinated visualization [132]. By coordinating different visualizations of the same dataset, users are able to combine information from each visualization to gain insight into underlying complex relationships in the data.

2.5 Visualization of Interpersonal Communication

Virtual human (VH) experiences are characterized by significant interpersonal communication. Thus any AAR of a VH experience must focus on interpersonal communication. Interpersonal communication is both verbal and nonverbal. Colloquially referred to as body language, nonverbal (or implicit) behavior refers to unplanned, non-deliberate, subconscious actions people use to communicate a wide variety of feelings and attitudes [34, 78, 124]. For example, Mehrabian [123, 124] has shown that people subconsciously use posture, gaze, and distance to express liking and dominance. Some researchers estimate that nonverbal behavior expresses as much as 93% of our feelings and attitudes [124] (the remaining 7% is encoded in verbal behavior). Thus, nonverbal behavior provides a window into a person's feelings and attitudes.

Of particular relevance for this dissertation is understanding interpersonal communication through visualization. IPSViz and IPSViz^N use temporal visualization to provide insight into the verbal and nonverbal communication between a human and virtual human. This communication is both verbal and nonverbal. Thus, this section reviews previous uses of visualization to understand and enhance verbal and nonverbal communication.

2.5.1 Textual Communication

Visualizations of textual communication (e.g., e-mail, text messages, and newsgroups) enable identifying patterns of communication, relationships among communicators, and themes and topics in the messages communicated. Themail [187] plots the major themes in a series of e-mails against the months of the year. More frequent, distinctive themes are rendered larger. At a glance, users can determine the themes of e-mail discussions, and how discussions differ from person to person. PeopleGarden [198] uses a growing flower metaphor to describe posting behavior on a message board. The height and number of petals on the flower correlate to time as a member of the message of board and posts, respectively. Aggregating flowers creates a PeopleGarden, which visualizes a message board's overall activity. Thread Arcs [93] depicts e-mail threads as chronologically sorted nodes with arcs flowing to replies. Filters color nodes and arcs to highlight characteristics of the e-mail thread (Who is participating in the conversation? Is the thread a business conversation or personal conversation?). Netscan [168] also visualizes e-mail threads. A tree depicts the hierarchical relationship between e-mails and replies, allowing a quick review of which e-mails led to the most (and least) discussion. A directed graph depicts which thread participants (nodes) reply to each other (arrows), allowing a quick determination of who communicates with who. A 3D bar chart renders each thread participant on the vertical, and days on the horizontal. For each participant-day pair, a bar is rendered depicting the number of posts the user made that day. This highlights temporal patterns in user interaction with a thread. Babble [52] depicts real-time participation in a chat room using a circle metaphor. Circles represent chat rooms

and dots in each circle represent participants. The closer dots are to the center of the circle, the more that participant is contributing to the conversation. This creates a visual representation of gathering for a conversation, much like in a real-world conversation. Those in other chat rooms are depicted as dots outside the circle. Thus, Babble depicts presence and level-of-activity in a discussion. Chat Circles [186] uses circles to represent users in a chat room. The color of the circle is used to represent identity, the brightness and size of the circle indicates activity (recent posting), and the proximity of circles enables discovery of communities of users discussing similar topics (much like small groups form at parties). Users can also review the Conversation Landscape, a history of each user's postings to discover who is dominant in a conversation, who users interact with, and whether a user's posts occurred near the user or far away. IPSViz and IPSViz^N provide similar visualizations of conversation history, topics and activity level.

2.5.2 Spoken Communication

As spoken communication is fundamentally different than textual communication, different visualization techniques are appropriate for spoken communication. These techniques rely on audio of the conversation participants and/or speech recognition to generate visualizations. Often the purpose of the visualizations is to show conversation participants a *social mirror*, [22], a reflection of hidden aspects of their social interaction including. The Conversation Clock [22] depicts the history of everyone's speech to highlight "turn taking, interruption, conversational dominance, silence, agreement, aural back-channels, mimicry, time spans, rhythm and flow." Other systems depict the major topics of the conversation [23] and the role, participation, and dominance of participants [11, 46]. Alallah et al. [5] create visualizations of face-to-face meetings to highlight the decision-making process. A timeline demonstrates when each participants speaks, when turns change, which decision each participant argues for (or against), and the final vote and outcome of the decision making process. Previous work has highlighted that real-time exposure to a social mirror can change one's interactions with others in a face-to-face

conversation [46]. IPSViz, IPSViz^N, and VSP have a similar aim. They present one's communication with a virtual human to encourage self-reflection and self-directed change in interpersonal communication skills.

2.5.3 Nonverbal Communication

Nonverbal communication has seen less interest from the visualization community, likely because it is more difficult to capture and represent nonverbal behavior for processing and visualization. The most prominent exploration of nonverbal behavior has been in visualizing eye and head tracking data. However, these visualizations have not been used to understand interpersonal communication inasmuch as understand general interaction with interfaces. For example, visualizations of eye gaze allow website designers to evaluate where users focus when looking at a web page [150]. Another work looks at how head gaze differs in virtual environments to determine what aspects of a virtual environment create breaks in presence [171].

Gaze visualizations take many forms. Fixation maps [196] aggregate fixation points - places where user gaze dwells - onto a surface. Higher areas on the surface represent where users focus more in a stimuli image. This surface can be modulated with the stimuli image to call attention to areas of low and high gaze (e.g., by brightening areas of high gaze). Ramloll et al. [148] create heat maps from fixation points to understand gaze on 3D objects. Ovals have been used to visualize clusters of fixation points and regions-of-interest - places where the users gaze (but do not necessarily dwell) [160]. Time plots are 2D static graphs depicting gaze paths, the paths between fixation points. If fixation points are represented as coordinates (x,y) in a stimuli image, a time plot ignores one dimension (say x) and plots the other (y) against time [143]. Another approach to reviewing paths between gaze points is by playing back a sequence of gaze points as a movie [20].

Some projects provide abstract visualizations of nonverbal behavior and their underlying affective meaning. VirtualWave [191] maps user head motion and breathing behavior to a series of 3d blocks arranged on a plane. The blocks expand and contract

to convey the starting and stopping of speech (from breathing) and form pyramids to communicate nodding behavior. Text-based conversation systems often provide tools to replace the missing nonverbal communication channel. While most conversation systems use human-like avatars to present nonverbal communication [32, 73, 116], more abstract visualizations have also been used. Some text-based systems display images of facial expressions to represent an overall emotion from a conversation partner [26, 107]. Others map affect to distinct colors [108], font properties (e.g., bold, italics, type, color, size) [91], and text animation [175].

2.6 IPSViz, IPSViz^N, and VSP

This dissertation builds on previous work by exploring the impact of after-action review for virtual human experiences. Three AAR systems are presented:

1. The interpersonal scenario visualizer (IPSViz) generates visualizations of the verbal and nonverbal communication between a human and a virtual human. Using IPSViz to conduct an AAR of one's interaction with a VH encourages reflection on one's interpersonal skills and self-directed change.

2. The Virtual Social-Perspective Taking system (VSP) leverages virtual experiences to enable reliving one's interaction with a VH through the VH's eyes. VSP allows users to reflect on the perspectives of others. Lessons learned through perspective-taking in VSP elicits changes in behavior in future interactions with real humans.

3. IPSViz^N presents aggregate visualizations of interpersonal communication among large groups of VH experiences. This enables users to identify trends in outliers in interpersonal communication among VH experiences and VH users.

Before discussing these systems, the next chapter reviews my work identifying the similarities and differences between interactions with a virtual human and interactions with a real human. By comparing interactions with a virtual human to those with a real human, this work identifies the verbal and nonverbal characteristics of VH experiences that must be incorporated into any AAR system for VH experiences.

CHAPTER 3
COMPARING INTERPERSONAL INTERACTIONS WITH A VIRTUAL HUMAN TO THOSE WITH A REAL HUMAN

This chapter describes two studies aimed at identifying the similarities and differences between interpersonal interactions with a virtual human and those with a real human. The first study was published in the proceedings of the IEEE conference on Virtual Reality 2006 [145] and an extended version discussing both studies was published in IEEE Transactions on Visualization and Computer Graphics [144].

Personal contributions: I designed and conducted both user studies, and analyzed the study data. I also was a key contributor to the development of the virtual human (VH) system used in the studies.

Collaborators: Kyle Johnsen and Robert Dickerson were also key VH system developers and helped with running the studies. Cyrus Harrison, Jonathan Jackson, and Min Shin provided tracking systems that enhanced the immersiveness and natural interactivity of the VH system. Marc Cohen, Margaret Duerson, Rebecca Pauley, Amy Stevens, Peggy Wagner, and D. Scott Lind provided access to study participants (medical students) as well as medical education expertise in the form of: content for the VH scenario, advice on study design, and ratings of participants in the virtual and real human interactions.

Relevance to Thesis: This chapter relates back to the thesis statement in two ways:

1. By establishing similarities between VH interactions and real human (RH) interactions on domain-specific interpersonal skills, this chapter motivates exploring after-action reviews (AARs) for VH interactions. AARs are effective pedagogical tools in RH interactions. The similarities between RH and VH interactions implies AAR might also be an effective pedagogical tool for VH interactions.

2. In the process of establishing similarities and differences between RH and VH interactions, this chapter identifies specific domain-oriented interpersonal skills that are exhibited in the RH and VH interactions. These are the skills that must be captured, processed, and displayed for after-action review.

3.1 Introduction

The Virtual Patient system [44, 87, 88] allows medical students to practice a difficult interpersonal situation - the medical interview - through interaction with a virtual human. Just as flight simulators help pilots improve flight skills, *interpersonal simulators* like the Virtual Patient system have the potential to help users improve *interpersonal communication skills*. This paper explores the potential of interpersonal simulators by comparing interactions with a virtual patient, *a virtual human that simulates a patient*, to interactions with a standardized patient, *a real human that simulates a patient*.

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

This paper describes two studies that compare standardized patient interactions to virtual patient interactions. Participants were medical students who interviewed either a) a standardized patient trained to simulate the symptoms of appendicitis (Figure 3-1 - left) or b) a virtual human programmed to simulate the same symptoms (Figure 3-1 - right). The interactions were then compared on the content of the interview, the behavior of participants, and the authenticity of the interaction.

Study I ($n = 24$, where n is the number of participants), presented at IEEE Virtual Reality 2006 [145], found that interactions with the standardized patient and virtual human were similar on gathering critical information from the patient and other content measures. Subtle differences were found on behaviors related to rapport with the patient. Participants appeared less engaged and insincere with the virtual human. Differences on rapport-building behaviors stemmed from the virtual human's limited

expressiveness. Ultimately, Study I was limited because it did not sufficiently characterize participant behavior. Study I highlighted the need to develop new measures of behavior in interpersonal interactions.

Building on Study I, **Study II** ($n = 58$) sought to 1) further characterize how behavior changes with virtual humans using new measures and 2) strengthen the main findings of Study I. Study II differed from Study I in the following ways:

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

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

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

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

3.2 Related Work

Although little work directly compares real and virtual interpersonal scenarios, researchers have compared other virtual environments to their real counterparts. In the psychology domain, Emmelkamp et al. [51] compared the reactions of acrophobes (persons with a fear of heights) in virtual and real environments. The authors found that exposure therapy in the virtual environment was as effective as therapy in the real one. Rothbaum et al. [158] found similar results for treating the fear of flying. Experiencing a virtual airplane was just as effective as experiencing a real one in reducing flying anxiety.

Others have looked at human perception of real and virtual stimuli. Billger [24] examined the perception of color in virtual and real environments. Wuillemin et al. [197] looked at the perception of virtual and real spheres presented visually and with haptics. Virtual spheres presented visually were perceived as larger than real spheres of the same size.

Heldal et al. [77] studied collaboration in real and virtual environments. Participants collaborated on building a Rubik's cube real or shared virtual environments. Performance in symmetric immersive environments (e.g., both participants collaborating through an immersive projection system) approached performance in real environments. Performance in asymmetric environments (e.g., HMD vs. immersive projection) was poorer.

Slater et al. [166] looked at the behavior of small groups in real and virtual environments. Participants viewed immersed peers as leaders in the virtual scenario but not in the real one. Group accord was higher in the real environment.

Usoh et al. [183] examined participant responses on presence questionnaires after experiencing a real environment or a similar virtual environment. Participants indicated they felt just as present in the virtual environment as in the real environment. This surprising result shows that subjective questionnaires should not be used to compare different environments.

3.3 Virtual Patient System

The Virtual Patient system [87] (Figure 3-2) allows practice of medical interview skills. Students can gather the key facts of a patient's condition and arrive at a differential diagnosis, a list of conditions the patient may have. They also can practice communicating clearly with the patient and addressing their fears. These communication skills do not just improve with clinical experience. They should be taught and practiced [103].

As part of the studies, the system was installed in a real medical exam room. A virtual exam room was projected on a wall. DIANA, a virtual human with severe stomach pain, was in the virtual room. DIANA's appearance and responses are modeled after a real standardized patient, Maria, trained to exhibit severe stomach pain. Modeling DIANA after Maria allowed participants to interact with similar patients in the real and virtual experiences. Another virtual human, VIC, served as an instructor that tutors students on how to interact with DIANA. Commercial speech recognition software and a simple algorithm for parsing utterances [44] enabled talking to VIC and DIANA naturally within the scope of the scenario. The student's hand was tracked, allowing DIANA's pain to be localized with pointing gestures. The student's head was also tracked to render the scene from her perspective and allow the virtual human to maintain eye contact.

3.4 Study I: Design

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

3.4.1 Measures

3.4.1.1 Eliciting critical information

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

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

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

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

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

3.4.1.2 Interaction behavior

Interactions were examined for behavioral differences between the two groups. Oviatt [134] observed spontaneous disfluencies (false starts, hesitations, etc.) occur less in machine-human interaction than in human-human interaction. Therefore, interactions were assessed on the conversation flow. Conversation flow was graded by counting the number of confirmatory words, like "ok" and "mmhmm," used in the interview. Such phrases are

often used when a person understands what the other is saying and wants to continue with the next topic. The expert observers also noted qualitative differences in conversation flow.

The interactions were also analyzed for empathetic behavior. Empathizing with the patient is a key component of building rapport. Empathy lets the patient know the doctor understands her situation [35]. Empathetic behavior is also an indicator of the participant's emotional involvement in the interaction. Participants' empathetic actions (e.g., saying "I know it hurts," acknowledging the patient's fears, etc.) were counted.

3.4.1.3 Perceptions of the interaction

The Maastricht Assessment of the Simulated Patient (MaSP) [195] is a validated survey used to evaluate standardized patients. To gather perceptions of the virtual and real interactions, participants filled out a modified MaSP focusing on authenticity and behavior. Questions on the MaSP include whether the patient is challenging/testing, the patient maintains appropriate eye contact and the simulated patient could be a real patient.

3.4.2 Participant Background

Group SP (*n*=8): Eight 2nd-year medical students (four male and four female) from the University of Florida interviewed the standardized patient (SP). On average, this group interviewed sixteen SPs prior to this study.

Group VH (*n*=16): Nine medical (four 1st-years, one 2nd-year, four 3rd-years) and seven physician-assistant students (four 1st-years, two 3rd-years, one 4th-year) from the Medical College of Georgia interviewed the virtual human. Seven were male, eight were female, and one did not specify a gender. On average, this group interviewed four SPs prior to this study.

3.4.3 Procedure

Figure 3-3 summarizes the Study I procedure.

3.4.3.1 Pre-experience

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

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

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

3.4.3.2 Experience

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

Group SP: Before the experience, participants waited outside the medical exam room door. When the standardized patient was ready, the words "You may now start the station" were played from an overhead speaker. The students then entered the examination room. The standardized patient was inside, lying on an examination bed. The standardized patient was in character as soon as the participant entered the room, and the interview began immediately upon entering. Participants were given up to ten minutes to conduct the interview. After eight minutes, a bell was played to warn participants that two minutes remained in the interview. After ten minutes, the following words were played from the overhead speaker: "Time is up. Please leave the station." As the participants had all experienced standardized patient interviews at the testing facility before, they were familiar with this procedure and the audio cues. The audio cues were scheduled on a strict timer. It should be noted that the amount of time spent interviewing

the standardized patient varied from participant to participant. Participants who finished early were allowed to leave the room and move on to the post-experience surveys.

Group VH: As in Group SP, participants waited outside the door of a medical exam room at the beginning of the experience. When the virtual human was ready, the study staff instructed participants to enter the room. Upon entering the room, participants sat in a chair and faced the projection of the virtual exam room. Figure 3-4 shows the virtual scene presented by the system. The virtual instructor, VIC, stood in the background and the virtual patient, DIANA, lay on the examination bed in the foreground.

Participants were instructed to say "hello" to VIC to begin the interaction. VIC responded by guiding participants through a short tutorial on interacting with the system. After the tutorial, VIC told the participant she had ten minutes to complete the interview. VIC then left the room so that the participant and DIANA could have privacy. The ten minute timer began as soon as VIC left. At the 8-minute mark, VIC informed the participant that two minutes remained over the system speaker (without reentering the room). After ten minutes, VIC returned to the room and ended the interaction. He then asked the participant for a diagnosis. After the participant stated their diagnosis, VIC thanked the participant and asked her to leave the room. As in Group SP, the amount of time spent interviewing the virtual human varied from participant to participant. Participants who finished early were allowed to leave the room and move on to the post-experience surveys.

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

3.4.3.3 Post-experience

Group SP: Participants related their perceptions of the standardized patient by filling out the MaSP survey (Section 3.4.1).

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

3.5 Study I: Results and Analysis

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

3.5.1 Statistical Analysis and Nomenclature

Throughout this section, a two-tailed Student's T-Test is used to test for significant differences ($\alpha < 0.05$). Note that items where differences were not found are not guarantees of similarity. Instead, the term similarity in this article denotes an inability to show statistically significant differences and a reasonable closeness in the mean and standard deviation. Statistical equivalence tests are gaining acceptance [135], and we plan to use them in future work.

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

3.5.2 Content Measures

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

3.5.2.1 Eliciting critical information

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

To assess the variability between the expert's observations, the total score on the critical information metric was correlated pair-wise across observers. The lowest Pearson correlation between the expert observers was $r = 0.86$. All correlations were significant at $p = 0.006$ or lower. This significant, large positive correlation indicates there was little inter-observer variation on the critical information metric. Observer scores were combined by averaging due to the low inter-observer variation. Mean values (M_{SP} and M_{VP}) represent the fraction of participants that elicited a critical piece of information.

According to the expert observers, participants asked the virtual human and standardized patient the same critical questions. As the virtual human's responses were based on the standardized patient's responses, the answers to the critical questions were also similar in both groups. No difference was found for both easily discussed information ("The pain is sharp and stabbing", $M_{SP} = 1 \pm 0$, $M_{VH} = 0.8 \pm 0.35$, $p = 0.12$) and sensitive information ("I am sexually active", $M_{SP} = 0.54 \pm 0.5$, $M_{VH} = 0.45 \pm 0.52$, $p = 0.72$). Final scores on eliciting the 12 critical items ($M_{SP} = 6.3 \pm 1.7$, $M_{VH} = 5.5 \pm 2.1$, $p = 0.37$) and

the fraction of students who received a passing grade ($M_{SP} = 0.5 \pm 0.54$, $M_{VH} = 0.36 \pm 0.5$, $p = 0.58$) were also similar. This suggests that a virtual human can sufficiently perform the role of a real person in a constrained, information exchange scenario.

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

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

It should be noted that the virtual human is the only true objective grader. This is because the virtual human graded participants by logging whatever information she revealed to them. The virtual human cannot take into account other factors when grading. In contrast to the standardized patient's grades, the virtual human's grades tend to match

the expert grades closely. For example, the standardized patient's grades on sexual history differed by 34% from the expert grades, but the virtual human's grades on sexual history differed from the experts by only 1%. The virtual human's ability to grade similarly to a panel of medical experts is a clear advantage over the subjective grading of standardized patients.

3.5.2.2 Educational goals

The similarities on content show that interactions with a virtual human can meet the educational goals of interactions with a real human. Participants in the virtual interaction were able to practice asking a patient questions, a key aspect of the medical interview. Furthermore, participants rated the virtual and real interactions' educational value similarly (Figure 3-5 - 1=Complete Disagreement, 5=Complete Agreement). One student said, "I thought it was really interesting, it was challenging and it was good to refresh my memory on a lot of communication and interviewing skills." Another student noted that the system allows one to practice the process of interviewing a patient without feeling nervous: "It was a lot less pressure than a real person, even a standardized patient. In there with the virtual patient, I wasn't worried about looking natural and confident ... looking natural to the real patient. I was out there taking time trying to figure out what's wrong with the patient." The virtual scenario was a valuable educational experience.

3.5.2.3 Empathy

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

Empathy encourages patients to share information. By expressing empathy, participants were working towards their task of eliciting critical information from the

patient. Also, the use of empathy is a sign that participants tried to engage the virtual human emotionally. This is encouraging, considering the virtual human is not a real person.

3.5.3 Behavior

3.5.3.1 Empathy

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

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

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

From an evaluation and training standpoint, these students gamed the system. They knew they could behave improperly with the virtual human without being penalized for it. For this system to be effective, it must be able to detect when students game the system and respond appropriately. The virtual human should make a comment or change

her mood to make it obvious to the participant that their lack of sincerity is improper. Making the virtual human more sensitive to improper behavior will encourage students to be more sincere.

3.5.3.2 Conversational behavior

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

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

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

Conversation flow: The flow of the conversation was also unusual. Normally, people use confirmatory phrases to regulate the conversation flow. Confirmatory phrases are short, one word acknowledgements ("Yeah", "uh-huh," etc.) or repetitions of what was just said. For example, the standardized patient might say, "My stomach hurts a lot."

The participant's response would be "OK. Your stomach hurts. Can you show me where the pain is?" Confirmatory phrases were used throughout standardized patient interviews to confirm what the patient said and signal the start of another question. Far fewer confirmatory phrases were used with the virtual human ($M_{SP} = 20 \pm 4.7$, $M_{VH} = 3.5 \pm 4.1$, $p = 6E - 5$).

3.5.4 Expressiveness: Virtual Human vs. Standardized Patient

3.5.4.1 Differences

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

The virtual human was much less expressive. Her voice had a regular volume and tone. Her face did not convey the right level of pain. She occasionally shifted her body or moved her hands, but her facial expressions did not change accordingly. Besides looking at the participant, the virtual human used no other explicit behaviors to indicate listening. Occasional delays in speech recognition produced delays in the virtual human's responses. Participants often interpreted this to mean the virtual human was not as engaged in the conversation. Feedback on the MaSP showed the standardized patient communicated how she felt better than the virtual human and appeared to be a better listener (Figure 3-6).

Previous results did not indicate the virtual human's lack of vocal expressiveness was a major deficiency of the system [88], and no significant difference was found between synthesized speech and more realistic recorded speech [43]. Therefore, no effort was put into improving DIANA's voice. However, this comparison was between different speech modes of the virtual system. When compared against the voice of the standardized

patient, the lack of professional voice quality impacted results. This study shows that professional voice talent is necessary.

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

3.5.4.2 Similarities

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

3.5.4.3 Effects of expressiveness

We hypothesize that the differences in patient expressiveness affected several measures. Global measures of authenticity (Section 3.5.5) and impressions of the interaction overall ($M_{SP} = 9.5 \pm 0.53$, $M_{VH} = 6.6 \pm 2.0$ on a scale of 1 to 10, $p = 1E - 4$) were lower in Group VH. Furthermore, although participants asked the virtual human and standardized patient similar questions, some behavior (Section 3.5.3) was different with the virtual human. Group VH asked questions in a more direct, rapid-fire fashion, and changes to conversation flow were observed. Empathy was expressed, but the empathy was not as sincere as that seen in the real scenario. Participants suggested that the virtual human be more expressive: "I would suggest to have more emotions into them. Maybe if there was more feelings, more emotional expression." The effectiveness of virtual humans are strongly impacted by their expressiveness.

3.5.5 Authenticity

As standardized patients are simulations of real patients, it is useful to compare standardized patients to real patients to judge their authenticity. The MaSP (Section 3.4.1.3) is a standardized survey filled out by medical students to assess standardized patient authenticity from a local and global perspective. Global measures look at overall impressions of the interaction, like whether the standardized patient acts like a real patient. Local measures look at specific components of the interaction, like whether the patient expressed pain realistically. Participants filled out the MaSP after interacting with the virtual human and standardized patient.

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

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

Given the differences on behavior and expressiveness, it is surprising that the virtual and real interaction were considered similar on local authenticity measures. One would expect *any* real interaction to always be considered more authentic than its virtual counterpart. We hypothesize the real and virtual interactions were similarly authentic

on local measures because participants applied different standards when rating local authenticity. Upon examining debriefing comments, it became clear Group VH evaluated the 'humanness' of the virtual human, whereas Group SP judged the accuracy of the standardized patient to a real patient. This is similar to Usoh et al.'s [183] conclusion that people apply different standards to real and virtual environments on presence questionnaires.

3.5.6 Post-Study Reflections

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

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

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

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

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

Virtual human voice fidelity: As part of a separate study, some Group VH participants spoke to a virtual human with a computer-generated voice, and others spoke

to a virtual human with pre-recorded real speech. No significant difference was found between the text-to-speech and real-speech conditions on all measures [43]. Therefore, the two groups were combined together in our analysis. However, as virtual human expressiveness has been identified as affecting results, recorded professional talent should always be used.

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

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

Study I's results also provide evidence that experience transfers from the real to the virtual interaction. The fact that students asked the same questions in both real and virtual interactions implies that Group VH brought their experiences with them into the virtual interview. Furthermore, in [89], a strong correlation was found in interaction skills between SP and VH interviews. Students who do well in VH interactions also do well in SP interactions. Likewise students who do poorly in VH interactions do poorly in SP interactions. This could not be possible unless experience transfers from the real to virtual.

3.5.7 Summary

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

3.6 Study II: Motivation and Design

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

- New behavioral measures were added to characterize rapport with the virtual human and standardized patient.
- All participants were 2nd-year medical students from the University of Florida. Restricting participants to the same institution and year of study controlled for differences in training and skill level.
- The virtual human used more expressive, pre-recorded speech. This guaranteed the voice fidelity of the virtual human was the same for all participants.
- The system was integrated into a patient communication course. The course gave access to a larger sample of the general medical student population. Instead of using volunteer participants, medical students were randomly selected to interact with the virtual human as part of their coursework. Participation was not compulsory.

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

3.6.1 Measures

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

3.6.1.1 Content

As in Study I, experts observers noted if participants asked the patient about twelve critical pieces of information needed to reach a diagnosis (Section 3.4.1). Reaching a diagnosis may be easier with a detailed picture of a patient's medical history. Therefore, participants were graded on whether they elicited information on five history categories: social history, family history, history of present illness, medical history and review of systems. Patient history information is not critical to reaching a diagnosis, but it can be helpful in narrowing down the list of topics to ask about.

3.6.1.2 Rapport-building behavior

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

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

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

Nonverbal communication: People use nonverbal behavior, often subconsciously, to communicate attitudes and feelings. For example, sustained eye contact, forward body lean and proper head nodding communicate attentiveness and an overall positive attitude [123]. Experts graded participants on nonverbal communication because it contributes significantly to rapport with the patient.

3.6.2 Procedure

Study I's procedure was modified to match the patient communication course procedure. Normally, students arrive at a medical education facility at a pre-determined hour. Each student is assigned two medical exam rooms where standardized patients are waiting. As soon as overhead speakers play "You may now start the station", the students simultaneously enter their first assigned room. The participants interact with the standardized patient for up to ten minutes. At the 8 minute mark, a warning bell is played to inform students that 2 minutes remain. When students complete their interview, they use any remaining time, plus a 2 minute break, to summarize what they learned about the patient and suggest a course of action. After the 2 minute break, the students repeat the process with their second standardized patient. Once a student is finished with the second interview, she is free to go.

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

Due to the strict course schedule, and a desire to make the virtual interaction like the real one, VIC was removed from the virtual interaction. As a result, the start, 2-minute warning and end sounds were all played from the facility's speakers. Also,

participants were not given a system tutorial. Initially, there was some concern that the loss of VIC's tutorial would lead to participant confusion. In fact, removing the tutorial made the experience more familiar to students. Students were able to start the interview immediately as they usually do in standardized patient interactions.

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

3.7 Study II: Results and Analysis

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

3.7.1 Inter-observer Reliability

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

Table 3-1 shows the pair-wise Pearson correlation of the observers on the three summary measures. Table entries contain an X where no correlation was found due to insufficient observer overlap. Most observers are reasonably correlated with each other ($r > 0.4$) and have a less than 5% ($\alpha < 0.05$) chance of being correlated due to chance. This implies the observers rated the interactions similarly.

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

3.7.2 Content Measures

3.7.2.1 Eliciting critical information

As in Study I, participants elicited the same critical information from the virtual human and standardized patient. Both groups were equally likely to elicit eleven of twelve critical pieces of information from the patient. Figure 3-7 compares overall performance on eliciting critical information in Study I and Study II. Not only was overall performance similar across groups, it was also similar across studies. The consistency on eliciting critical information across studies strengthens the assertion that content was similar in the real and virtual interactions.

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

3.7.2.2 Patient history

As an additional content measure, experts graded participants on their ability to elicit patient history information. Patient history provides more background that can help the student reach a diagnosis. Overall, participants elicited less patient history information from the virtual human. For example, Group VH was less likely to gather family medical

history ($M_{SP} = 0.5 \pm 0.48$, $M_{VH} = 0.22 \pm 0.38$, $p = 0.018$) and the history of the present illness ($M_{SP} = 0.85 \pm 0.26$, $M_{VH} = 0.63 \pm 0.32$, $p = 0.008$).

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

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

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

3.7.3 Rapport-Building Behavior

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

3.7.3.1 Process and etiquette

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

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

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

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

3.7.3.2 Empathy

As in Study I, participants in both groups responded empathetically to their patient ($M_{SP} = 0.79 \pm 0.29$, $M_{VH} = 0.69 \pm 0.42$, $p = 0.34$). However, the sincerity of the empathy was lower with the virtual human. Group VH's empathetic behavior remained robotic and disengaged in Study II. 27% of Group VH used spontaneous empathy versus 84% in Group SP ($p = 4.38E - 6$). A close to significant difference was found on the overall quality of the empathy on a scale of 1 to 4, ($M_{SP} = 2.7 \pm 0.85$, $M_{VH} = 2.3 \pm 0.84$, $p = 0.08$)

Surprisingly, specific, descriptive ratings of empathetic behavior in the virtual and real interactions were not different. For example, Figure 3-8 shows that both groups were rated similarly on descriptive scales like "good/bad," "weak/strong," and "active/passive." This is in stark contrast to the overall sense that empathy behavior was poorer with the virtual human. We hypothesize no differences were found on these scales because they are too subjective. The expert raters could not objectively rate abstract concepts like "weak/strong." As part of future work, we are exploring the use of objective behavioral measures to augment these subjective measures.

3.7.3.3 Nonverbal communication

Nonverbal communication is critical to rapport-building because it communicates a variety of feelings and attitudes. Group VH's nonverbal behavior expressed lower rapport with the virtual human (Figure 3-9 - 1=Very Poor, 4=Very Good). They used less forward body lean and nodded less. These behaviors were inappropriate because they are associated with lower interest and a poorer attitude [123]. Not surprisingly, Group VH appeared less attentive and had a less positive attitude with the virtual human (Figure 3-9).

It should be noted that expert ratings of participants' eye contact were similar with the virtual and real human. This was also seen in Study I, where participants indicated the virtual human and standardized patient maintained good eye contact. The virtual human constantly looked at the participant throughout the interview, influencing

participants to reciprocate and maintain eye contact. Future studies should incorporate an eye tracker to confirm this result and determine the amount of eye contact more accurately.

3.7.4 Expressiveness of the Virtual Human

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

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

3.7.5 Summary

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

3.8 Conclusions

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

content measures. Participants gathered the same information from the virtual human and standardized patient.

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

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

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

3.9 Future Work

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

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

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

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

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

Table 3-1. Pearson correlation between observers.

Pearson Correlation (r)

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

Significance (p)

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

Figure 3-1. Real interpersonal interaction (left) and equivalent virtual interpersonal interaction (right).

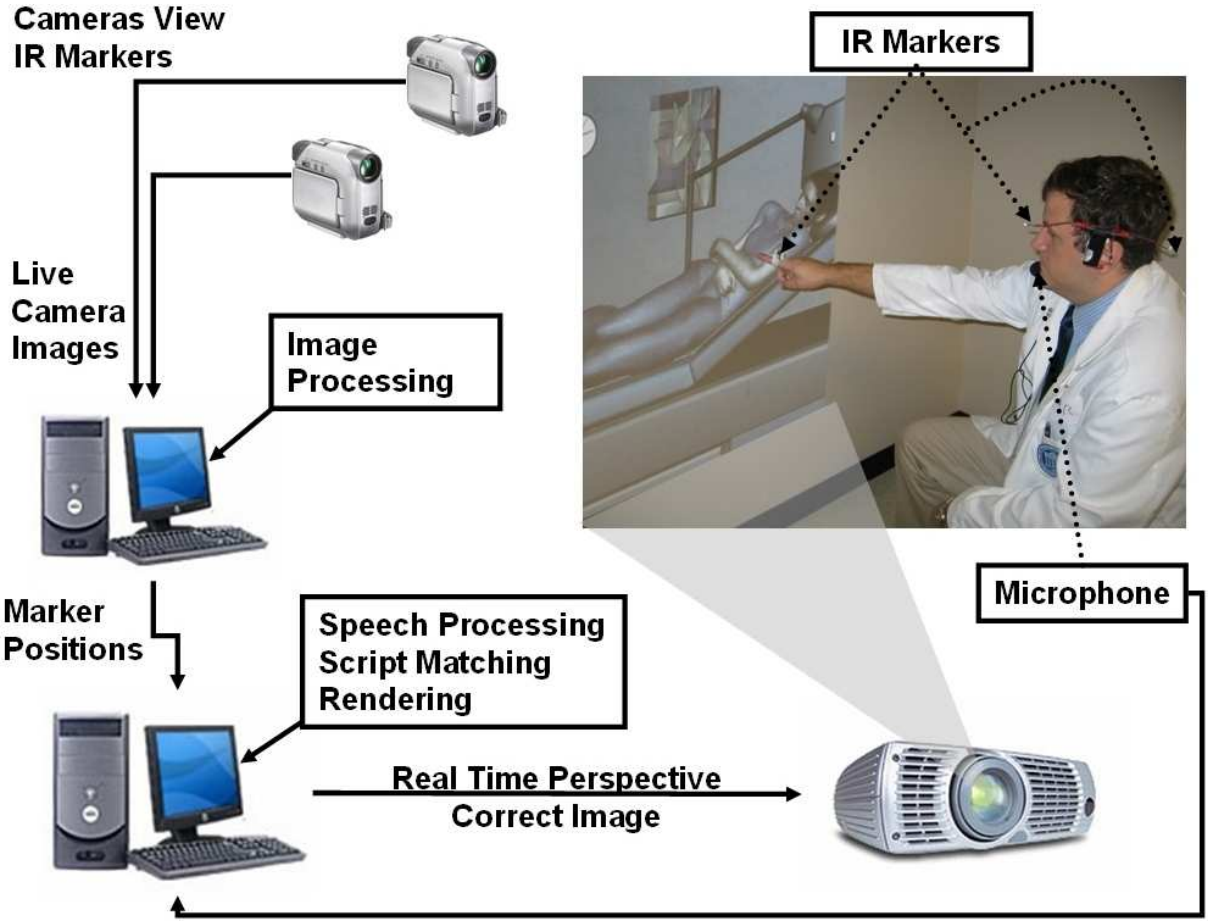


Figure 3-2. System overview

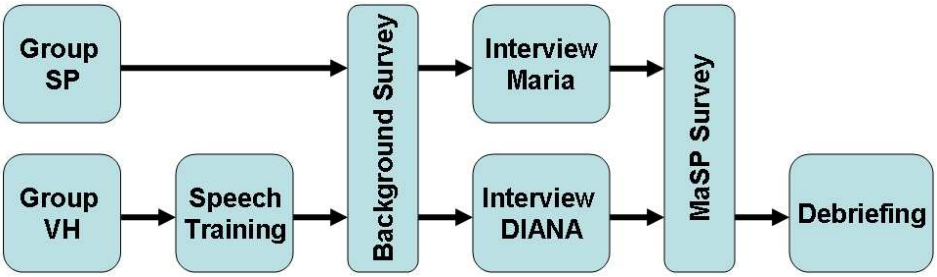


Figure 3-3. Study procedure for groups SP and VH

Figure 3-4. VIC (left) and DIANA (right) in the virtual exam room

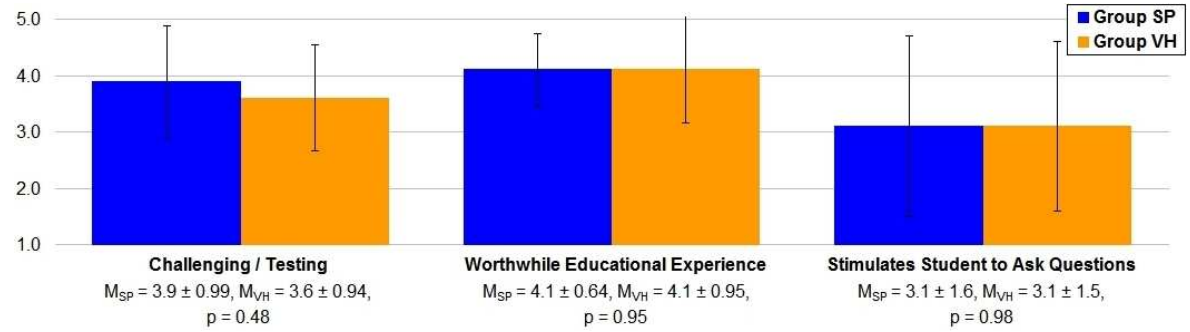


Figure 3-5. Perceived educational value of real and virtual interactions

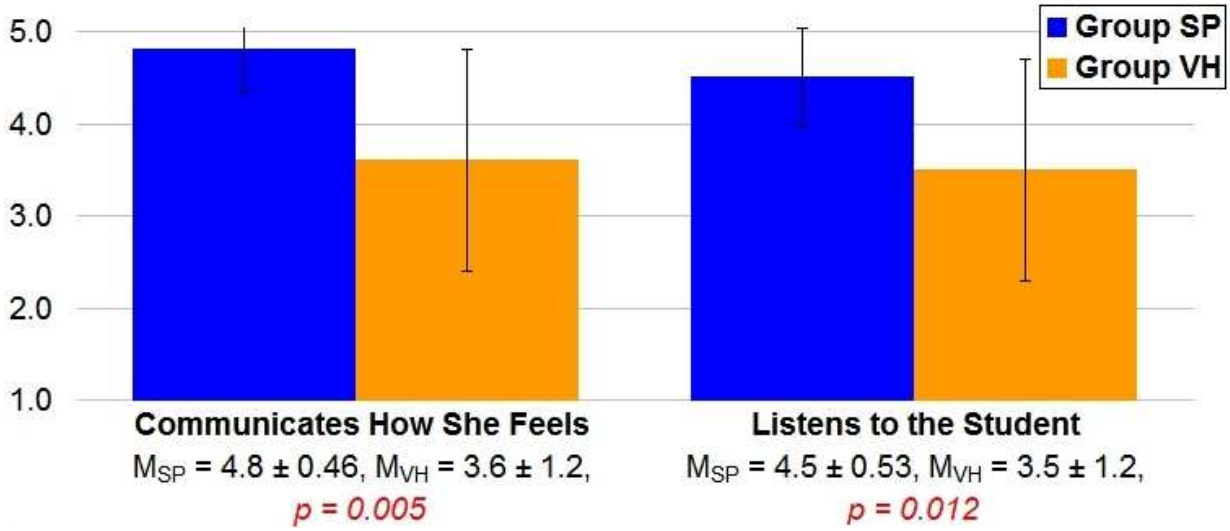


Figure 3-6. Perceived expressiveness of virtual human and standardized patient

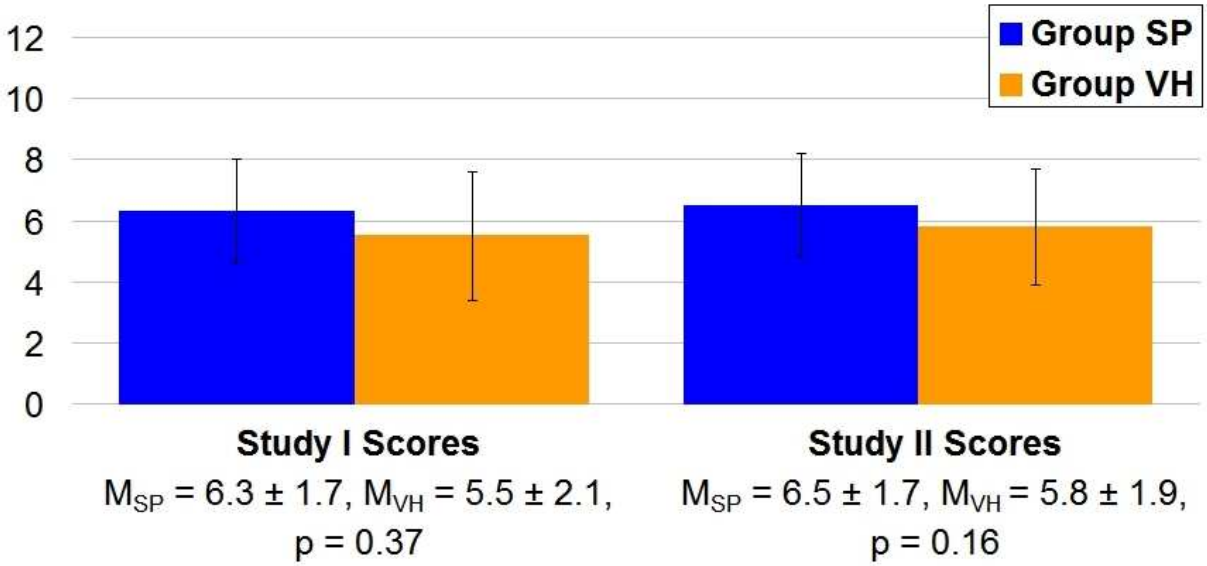


Figure 3-7. Eliciting information score for both groups and studies.

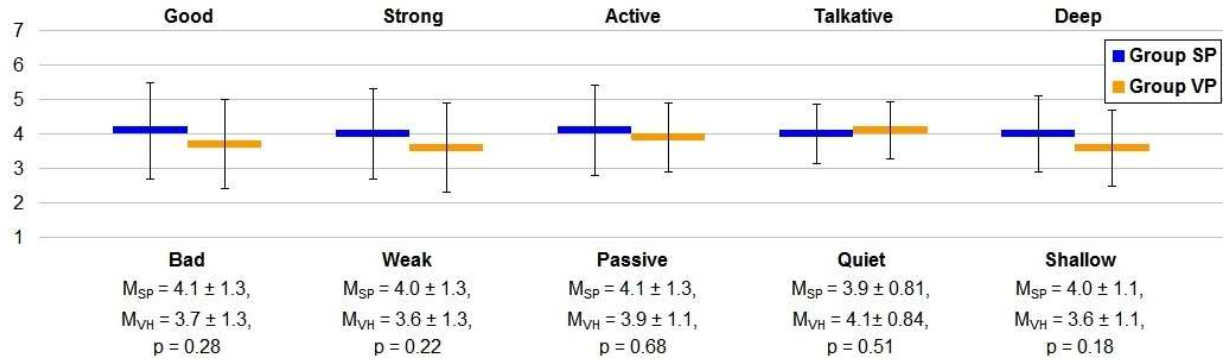


Figure 3-8. Expert ratings of empathy on descriptive scales

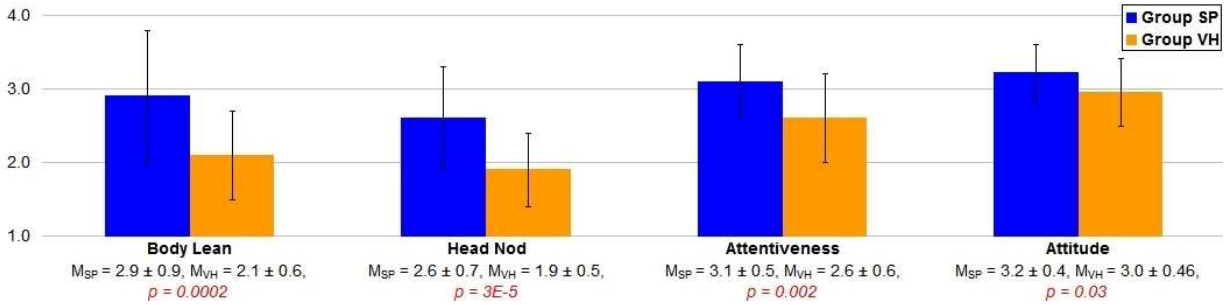


Figure 3-9. Nonverbal behavior, attitude, and attentiveness

CHAPTER 4
IPSVIZ: HUMAN-VIRTUAL HUMAN EXPERIENCES FOR SELF-REFLECTION AND SELF-DIRECTED CHANGE

This chapter describes, IPSViz, a desktop-based after-action review (AAR) system for interpersonal interactions with a virtual human (VH). IPSViz leverages visualization to enable review of interpersonal communication with a virtual human. A pilot study reviews the impact of IPSViz on users of VH experiences. This work was published in the proceedings of the IEEE conference on Virtual Reality 2008 [147].

Personal contributions: I developed IPSViz. I also instrumented the VH system with a logging framework to collect data on a user's communication with a VH. Lastly, I designed and conducted the user study and analyzed the study data.

Collaborators: Kyle Johnsen, Brent Rossen, and Aaron Kotranza assisted with the development of the VH system and provided help with running the studies. Joshua Horton developed the original "topic plot" visualization discussed later in the chapter. Rick Ferdig provided expertise on study design and an educational viewpoint on this work. D. Scott Lind and Adeline Deladisma provided access to study participants (medical students) as well as expertise on interpersonal communication in the VH scenario (the medical interview).

Relevance to thesis: This chapter describes IPSViz, one of the three AAR systems for H-VH experiences that were developed as part of this work. First, the chapter discusses the pipeline used throughout this dissertation to capture, process, and display human-virtual human interaction. Second, it demonstrates using this pipeline to generate visualizations of the interaction for after-action review of domain-specific interpersonal skills. Third, it shows that reviewing these visualizations elicits self-reflection, changes self-perceptions, and motivates change in domain-specific interpersonal skills. Lastly, it provides justification for two other types of analysis systems for human-virtual human experiences: reliving a VH interaction from the VH's perspective (VSP) and aggregate visualizations of groups of human-virtual human experiences (IPSViz^N).

4.1 Introduction

In previous work presented at IEEE VR 2008 [147], reviewing one's interaction with a virtual human was found to change self-perceptions of communication skills. The effect was strong enough that study participants claimed they would change behavior with real humans as a result of the AAR. Such self-directed change would be expected for interactions with real humans, but was surprising for interactions with virtual humans. State-of-the-art virtual humans can be similar to real humans, but they clearly do not meet the visual and behavioral realism of real humans [144]. Thus, it would be unexpected for users to review their interaction with a virtual human and better understand their interactions with real humans. This article builds on this unexpected result by exploring the relationship between the realism of the virtual human and the impacts of after-action review (AAR). This exploration led to two new discoveries: 1) an initial definition of the scope of impacts an AAR of an H-VH experience can have on a user, and 2) evidence that such impacts are attenuated, but not eliminated, by virtual humans with low interaction realism.

4.1.1 AAR of a Human-Virtual Human Interaction

Human-virtual human (H-VH) interactions are immersive virtual experiences that simulate social scenarios. Existing scenarios include military leadership [79], law enforcement [59], cultural competency [10, 40], medical interview and diagnosis [89, 115, 144], and interaction between children [159]. As training for H-H social situations requires after-action reviews (AAR), so does training with H-VH interactions.

After-action reviews play an important role in communication skills training. Communication skills training is a major component of business, military, and medical education. In business, military, medical education, and other fields where communication skills are crucial to success, instructors teach students communication skills using lectures, role-play, and situational immersion with instructor-observation. A critical educational component of these methods is the AAR. In AARs, students review their social interaction

and are evaluated either by themselves, their instructors, their peers, or a combination of these. The evaluation provides students with the feedback they need to improve their communication skills.

Recently, VHs have been applied to simulating and educating communication skills. Thus, we propose augmenting H-VH interactions with AAR to improve communication skills education with VHs. One of this paper's aims is to provide tools that support AARs of H-VH interactions. Such tools allow a user to interact with a VH, and then review their communication with the VH minutes afterward to learn how to improve future communication with real humans.

4.1.2 Interpersonal Scenario Visualizer (IPSViz)

The interpersonal scenario visualizer, or IPSViz, is an AAR tool for reviewing interactions between a human and a virtual human. IPSViz incorporates three design principles that enable users to gain insight from H-VH interactions:

1. An H-VH interaction is composed of social, temporal, and spatial characteristics. Thus, users should be able to explore these characteristics in IPSViz. IPSViz leverages visualization to explore these characteristics.

2. An H-VH interaction can be characterized as a set of signals. Thus, *interaction signals* are captured, logged, and processed to produce AAR visualizations.

3. An H-VH interaction is complex. Thus, users gain insight into this complexity by reviewing multiple visualizations, such as audio, video, text, and graphs.

To enable AAR for communication skills education, IPSViz generates visualizations of H-VH interactions. Novel visualizations are generated by leveraging the many interaction signals that are captured in an H-VH interaction. Given an H-VH interaction, AAR is facilitated through the following visualization types:

- **Spatial visualizations:** The H-VH interaction can be 3D-rendered from any perspective, including that of the conversation partner (the virtual camera is located at the VH's eyes). Students are able to perceive "what it was like to talk to them."

- **Temporal visualizations:** Events in the H-VH interaction are visualized with respect to an interaction timeline. Students are able to discern the temporal relationships (e.g., cause and effect) between conversation events.

- **Social visualizations:** Verbal and nonverbal behaviors are presented in log, graph, and 3D formats, allowing students to see how their verbal and nonverbal behavior affects the conversation.

4.1.3 Evaluating AAR for H-VH Experiences

The impacts of AAR with IPSViz were analyzed in a user study (N=27). Study participants were health professions students. The students interacted with a virtual human patient and then reviewed the interaction afterward in IPSViz. Before and after using IPSViz, students evaluated their interaction on interpersonal skills. The impact of IPSViz was measured by comparing participants' pre- and post-AAR evaluations of their communication skills.

The study found that *the H-VH experience does not end when the user leaves the virtual environment.* Rather, the impact of the virtual experience continues into the after-action review. Through AAR, self-reflection, changes in self-perceptions, and self-directed change are possible with human-virtual human experiences. Furthermore, human-virtual human experiences elicit the *most* impact on users when they 1) are coupled with after-action reviews to the experience, and 2) strive for high VH interaction realism.

4.2 Interpersonal Scenario Visualizer

4.2.1 Expanding AAR to H-VH Interactions

By expanding AAR to H-VH interactions, IPSViz builds upon the previous work in AAR. Unique to IPSViz is the focus on reviewing a social experience. Social-experience VEs have very different goals than spatially-driven VEs. For example, reviewing a user's posture and speech has vastly different meanings in a H-VH interaction than an architectural walkthrough VE. Thus, new approaches and techniques need to be applied to AARs for H-VH interaction. IPSViz's visualizations draw inspiration from the substantial body of research into interpersonal communication, medical communication, and psychology. In return, IPSViz adds to application and basic research by providing a significant value-add to simulating social situations for training and education.

4.2.2 Overview of IPSViz

IPSViz is based on representing a H-VH interaction as a set of signals. The signals include user speech, video, tracking data, and VH behavior. IPSViz processes and visualizes these interaction signals to enable AAR (Figure 4-3).

To generate visualizations of an interaction, the interaction is captured from a variety of sensors (Section 4.3). From a signal analysis perspective, capture is equivalent to sampling the interaction as if it were a set of continuous signals. These interaction signals characterize the interaction between a human and virtual human.

Before generating visualizations, interaction signals may undergo filtering and processing (Section 4.4). In both stages, a chain of digital filters is applied to one or more signals to derive new signals. Filtering and processing are separated into two stages as each solves a different problem. Filtering compensates for errors caused by sampling a continuous signal with real-world sensors (e.g., discretization error and noise). Processing manipulates and combines signals to provide new information about the interaction.

After filtering and processing, interaction signals are mapped to the visual (or other perceptual) domain to produce visualizations (Section 4.5). The visualizations allow users to gain new insight into H-VH communication.

4.2.3 A Human-Virtual Human Experience

The discussion of IPSViz is guided by a real-world H-VH experience - the interaction between a health professions (HP) student and a VH [88] (Figure 4-2). This experience was chosen to guide the discussion because HP students 1) take this interaction seriously [89], and 2) need review, evaluation, and feedback to improve their communication with patients [35, 157].

A typical interaction between a HP student and a VH begins with the VH complaining of a medical problem (e.g., pain). The student's goal is to determine what the problem is (diagnosis) and treat it. Effective diagnosis and treatment requires gathering accurate

information from the VH. The student gathers information from the VH by asking the VH questions and examining the VH.

During the interview, the VH will also ask the student questions to learn what is happening to her and why. The VH may also ask questions in the hope that the student can relieve her anxiety about the medical problem. Typical questions a VH can ask include "Do you know what I have?" and "Do you think this could be cancer?" Students should answer these questions carefully to relieve the VH's anxiety and build rapport with the VH.

4.3 Capturing Human-Virtual Human Communication

Communication is the means by which the VH and user reach their goals. Hence, logging communication is crucial to understanding the interaction. As communication is mediated by the system's input and output devices, system inputs and outputs are logged to capture H-VH communication.

4.3.1 System Inputs

Natural speech: Students wear a wireless microphone on their head. This enables talking to the VH using natural speech. Speech recognition software (Dragon Naturally Speaking 9 Pro) extracts the words spoken by the user from microphone input. Both the speech waveform and the speech recognition output are logged.

Head, hand and body lean tracking: Students wear a hat and glove outfitted with reflective markers. Also, markers are attached to the back of the student's chair. Combining head and chair tracking data enables computing approximate body lean. The markers are tracked by an optical, infrared tracking system (# cameras: 2 to 3, # objects: up to 4, DOF: 6, Update Rate: 60Hz, Latency: 100ms, Registration: $<$ 1cm, Jitter: 10mm). This allows the system to detect user presence, head gaze, pointing gestures, and chair motion. Detected events and the 3D positions of each marker are logged.

Video: Video of the interaction is also recorded for later review in IPSViz. Video is recorded because it is a standard practice for students and instructors to review videos of patient interactions.

4.3.2 System Outputs

Natural speech and animation: When a student speaks to the VH, the speech recognition software interprets her words. A keyword matching algorithm matches speech recognition output to questions in the VH's response database. If a match is found in the database, the VH executes a corresponding vocal and gesture response. The VH responds appropriately to the question 60 to 70% of the time. The VH's gesture and speech responses are logged.

Visual immersion: The interactions take place in a medical exam room or a mock exam room in a controlled laboratory. A projection display (NEC VT770 Projector or equivalent, 1024x768, 60 to 100 degree diagonal FOV depending on distance to display surface) or a head-mounted display (EMagin Z800, 800x600, 40 degree diagonal FOV) is used to show the user the VE. The VE is rendered at life-size, such that virtual objects appear to have the same proportions as they would if they were real. The VE is rendered in real-time (35 to 45Hz) using the OGRE 3D rendering engine.

Using the head-tracking inputs, the system renders the VE from the student's perspective. This allows the student to perceive the VH's gaze behavior accurately. The head tracking data also allows the VH to respond when the student enters the room. In the HMD condition, head-tracking enables looking around the virtual room. Life-size, user-perspective rendering and VH responsive behaviors create a highly immersive experience. 3D environment parameters, rendering parameters, and VH gaze are logged.

4.3.3 Summary of Captured Data

Capturing a ten-minute H-VH interaction produces on average 70 megabytes of data, stored into six separate files:

- **Transcript:** Timestamped list of events including the speech and gestures of the VH and student. As user speech and gestures are interpreted by imperfect speech and gesture recognition, this data can contain errors.
- **Student state:** Binary file that describes the state of the student over the course of the interaction. This is primarily tracking data.
- **VH state:** Similar to the student state log. It describes the state of the VH over the course of the interaction. VH state includes head pose and a variable indicating if the VH is talking.
- **Audio:** Audio recording of the student's speech. This is recorded by the wireless microphone worn by the student. Note that audio of the interaction is also recorded to the video log.
- **Video:** Video recording (360x240) of the VH and user. This is captured by a mini-DV camera.
- **System:** System and rendering parameters.

Note that this list of data is not meant to be exhaustive, nor completely representative for all VH applications. Rather, the types of data captured were chosen because of the importance of:

- verbal and nonverbal behavior in communication [78, 124],
- communication content and rapport-building in the medical interview [35, 167], and
- communication content and rapport-building in H-VH experiences [10, 36, 40, 79, 89, 115, 159].

4.4 Filtering and Processing

The captured interaction data is interpreted as a set of signals that can be filtered and processed. Filtering and processing is necessary to 1) compensate for sensor error, 2) combine signals to form new signals, and 3) extract signals embedded in other signals.

Sensor Error: Real-world sensors introduce errors into the signals they sample. Errors include noise, discretization error, information loss, and aliasing. For example, in optical trackers, these errors correspond to jitter from CCD noise and discrete pixels, data loss due to occlusion and limited tracking volumes, and aliasing when motion is faster than the Nyquist rate. Filtering cannot correct for these errors, but does compensate

for them to acceptable tolerances. Filtering is crucial because it prevents excessive propagation of sensor error into the visualizations that users review.

Combining signals: The measurement of body lean provides an example of combining signals to form new signals. One way to measure forward body lean is to measure the angle between the student's back and an up-vector (a vector perpendicular to the floor of the room). Unfortunately, it is difficult to track a student's back with an optical, marker-based tracker. Markers attached to the back could cause discomfort and be occluded when the student leans back in the chair.

To overcome these issues, an approximate body lean signal $L(t)$ is computed by combining head and chair tracking data. First, head and chair data is filtered to compensate for tracker jitter. Then the chair position is subtracted from the head position to compute a head-chair vector. The head-chair vector is a substitute for a more accurate body lean vector that runs along the student's spine. $L(t)$ is set to the angle between the head-chair vector and the up-vector.

$L(t)$ is a reasonable approximation of body lean because it allows identification of 1) when body lean changes and 2) the direction of the change (forward or backward). For example, the student represented in Figure 4-4 did not change his body lean except for leaning forward and then back around 3:12.

Embedded signals: The identification of discussion topics provides an example of extracting embedded signals. Topic signals characterize the discussion of a topic in an H-VH interaction. As the transcript log contains the speech content of the interview, topic signals are computed by filtering the transcript for words associated with a topic of interest. For example, if the topic is "family history," then the transcript is filtered for speech events that contain words like "family," "sister," and "mother." For all times associated with these speech events, the topic signal $F(t) = 1$, else $F(t) = 0$.

4.5 Spatial, Temporal, and Social Visualization

IPSViz generates visualizations to help health professions students evaluate interactions with VHs. Students evaluate themselves by asking certain kinds of questions about the interaction.

- *How much time* did I spend discussing the VH's symptoms?
- *Where* was the VH's pain?
- *When* was I empathetic to the VH?
- *Was there a moment when* my nonverbal behavior affected rapport with the VH negatively?
- Did I look *at* the VH *when* she was talking, or *elsewhere*?

As the italicized words above show, these questions are *spatial* and *temporal* in nature. The spatial questions focus on where objects are and how they are related to each other in the space of the 3D world. The temporal questions focus on when events happen and how long they happen for. Furthermore, these questions focus on how the student behaved *socially* with the virtual human. IPSViz generates visualizations that are spatial, temporal, and social to help students gain insight into their communication with the VH.

4.5.1 Spatial Visualization

IPSViz renders the space of the H-VH interaction using 3D models of the student, VH, and the mixed interaction environment (real room + virtual medical exam room). Using tracking data, the poses of the VH, user, and chair models are updated to reflect their motion during the interaction. This allows students to review their behavior in the space of the interaction environment.

Multiple viewpoints: Using a 3D representation of the interaction enables rendering the interaction from multiple viewpoints. This allows students to see what their behavior looked like to an external observer or to the VH (Figure 4-6A). Seeing the interaction through the VH's eyes is a powerful way of demonstrating to students how their nonverbal behavior is perceived by their patients.

Augmenting the environment: By using a 3D representation of the interaction, spatial information about the student's communication can be augmented to the environment. IPSViz demonstrates this by augmenting the 3D environment and VH models with a **gaze target** (Figure 4-6B). The gaze target is a texture that is projected wherever the user was looking during the interaction. The gaze target allows students to become aware of where their attention was actually focused, as opposed to where they thought it was focused.

4.5.2 Temporal Visualization

IPSViz allows users to explore an H-VH interaction temporally through nonlinear review and scalable timelines.

Nonlinear review: Students can play back the H-VH interaction nonlinearly. While in playback mode, the visualizations (video, 3D rendering, and timeline plots) are updated to present information relevant to the current playback time. The audio of the interaction is also played back so that students can hear themselves talk to the VH. Similar in principle to nonlinear video editors (e.g., Adobe Premiere), students select moments from a timeline to instantaneously play them back. This allows students to review the interaction in any order. The timeline is represented visually in the interface by a slider (Figure 4-1 - bottom). Additionally, events can be selected from a transcript (Figure 4-1 - right) to review them.

Scalable timelines: Interaction events and signals are depicted visually on timelines. This allows users to see the temporal relationships among events. Timelines are scalable. At the global time scale, users review the entire interaction at a glance. Local features are narrowed in on to review them in detail. This is similar in spirit to Shneiderman's mantra [161]: "Overview first, zoom and filter, then details-on-demand." The topic (Figure 4-5) and body lean plots (Figure 4-4) demonstrate the use of scalable timelines.

4.5.3 Social Visualization

IPSViz highlights social aspects of the interaction by visualizing verbal and nonverbal communication.

4.5.3.1 Verbal communication

Verbal communication is presented textually with a transcript (Figure 4-1 - right), graphically with a topic plot (Figure 4-5), and aurally by playing recorded audio of the interaction.

Transcript: The transcript is a text display of everything the VH and student said to each other. By reviewing the transcript, the student learns what information he did and did not learn from the VH. The transcript also serves as a table of contents into the interaction in that it helps users find important events to review.

Topic plot: The topic plot is an overview of the conversation. It filters the speech in the conversation down to a set of topics relevant to the scenario (e.g., symptoms or pain) and summarizes the use of these topics by plotting them on a timeline. Reviewing the topic plot allows students to see if they progressed from topic to topic logically, or if the flow of the conversation was confusing. They can also see if they forgot to discuss an important topic.

4.5.3.2 Nonverbal communication

Three types of nonverbal communication are presented in IPSViz, gaze, posture, and paralanguage.

Gaze: Gaze builds rapport by communicating attention, friendliness, and respect [8]. Hence, gaze is highlighted in IPSViz by rendering the gaze target and head pose of the 3d model of the student. In addition, rendering the scene from the VH's viewpoint allows students to review their head gaze behavior from the patient's perspective, and learn that patients notice head gaze behavior and know when they are not paying attention.

Posture: Posture is used in interpersonal interaction to communicate friendliness, interest, and social status [124]. The student can review their posture in IPSViz by

watching himself on video, looking at the body lean of the 3D model that represents them, or by looking for peaks and valleys on the body lean plot.

Paralanguage: Paralanguage is the set of non-speech vocal expressions, such as tone of voice, volume, and speech pauses [78]. Playing audio of the interaction allows review of paralanguage.

4.6 Study Design

A study was conducted to determine how AAR with IPSViz affects perceptions of an H-VH interaction. Participants (N=27) in the study were health professions (medical and physician assistant) students. Each participant conducted a medical interview with a VH and then reviewed the experience in IPSViz. Before and after using IPSViz, the participant graded her interaction on how she appeared to the VH (e.g., friendly or emotional), the emotional state of the VH (e.g., scared), and patient communication skills (information gathering, rapport, and procedure). Students were also interviewed to gather comments about what they learned from using IPSViz. This is a within subjects, repeated measures design with two levels (Pre-AAR and Post-AAR). This design enabled measuring how IPSViz changed participants' views of their interaction with the VH.

In addition, two separate studies were conducted concurrently to evaluate the effect of VH skin tone (light vs. dark) and system display type (head-mounted vs. projection display) on the H-VH interaction. VH skin tone and display type are between-subjects factors that were varied randomly among the N=27 participants. The between-subjects results from these studies will be reported in future publications. While no cross interactions between VH skin tone, display type, and AAR were expected, these factors were taken into account in the analysis as a precaution.

4.6.1 Participants

23 third-year medical and 4 first-year physician assistant students were recruited (15 male, 12 female). All participants were students at the Medical College of Georgia. Participants were paid $20.

4.6.2 Study Procedure

Figure 4-7 summarizes the study procedure.

Pre-virtual human interaction: Before interacting with the VHs, the participant signed a consent form and filled out a background survey. Participants in the HMD subgroup wore an HMD tracked with 6 DOF. Participants in the projector subgroup wore a hat tracked with 6 DOF. A wireless microphone was attached to the HMD and hat so that participants could talk to the VH. Participants spent two minutes training the speech recognition software.

Virtual human interactions: Participants conducted two interactions with VHs, a practice interaction and a main interaction. In the practice interaction, the VH was DIANA (Digital Animated Avatar), a 19-year-old female complaining of stomach pain. A study proctor began the practice with a tutorial on communicating with the VH. Then participants interviewed DIANA. The proctor remained in the room to assist the participant. The practice interaction was no more than five minutes long.

In the main interaction, the VH was EDNA (Elderly DIANA), a 55-year-old woman who found a mass in her breast. Participants were instructed to interview EDNA as if she were a real patient. Participants were told they would not receive any assistance and the proctor did not enter the room with the student. The main interaction was between five and ten minutes long.

Interviewing a patient who found a mass in her breast is a challenging situation for an HP student. We chose to amplify this challenge and increase the realism of the scenario by having the VH challenge the student at three separate points in the interview.

- **Challenge 1:** The VH said "I found a mass in my breast, and uh, I'm really worried about it."
- **Challenge 2:** The VH said: "Could this be cancer?"
- **Challenge 3:** The VH sneezed.

In a real medical interview, the first two challenges should elicit empathetic, comforting responses from the student, and the last challenge should elicit a socially appropriate response (e.g., "bless you" or "gesundheit"). Students reviewed these challenges in IPSViz to evaluate if they responded appropriately.

Pre-AAR survey: After the main interaction, students filled out surveys to gather their perceptions of their performance in the main VH interview. The surveys are discussed in Section 4.6.3.

IPSViz review sessions: Students conducted a practice AAR and a main AAR using IPSViz. In the practice AAR, students reviewed their practice VH interaction. A study proctor gave a five-minute tutorial on IPSViz, and then students were given a few minutes to become familiar with the IPSViz interface.

In the main AAR, students used IPSViz to review the main VH interaction. First, students reviewed the interaction on their own. They were encouraged to use IPSViz any way they wanted, including changing 3D viewpoints, reviewing plots, and skipping to important events. The study proctor did not interact with participants during this time. When the student was done, the study proctor conducted a short, guided review of the interaction. The guided review directed students to review their reactions to the three challenge statements (as described previously). The main AAR took approximately 10 minutes.

Post-AAR survey and debriefing: After using IPSViz, students filled out a survey to learn how perceptions of their performance changed due to the AAR. Finally, students were debriefed to collect qualitative data about the H-VH experiences and IPSViz.

4.6.3 Measures

Two types of data were collected from participants, background data (e.g., demographics or educational background) and self-evaluations of the H-VH interaction. The background survey was filled out before the VH practice interaction. The self-evaluation surveys were filled out before and after using IPSViz.

4.6.3.1 Friendly, natural, and emotional expression

In previous studies, some participants asked the VH questions in a stilted, emotionless, rapid-fire manner [144]. To assess if participants recognized this improper behavior during AAR, participants evaluated themselves on naturalness, expression of emotion, and friendliness. Five "friendliness" scales were used - pleasant, cruel, cold, unfriendly, and unlikeable (seven point Likert Scales, 1 = Not at all, 4 = Neutral, and 7 = Very). The friendliness measures have high internal consistency for real human interactions ($\alpha > 0.9$) [47].

4.6.3.2 State of the virtual human

In addition to helping patients medically, doctors need to help their patients emotionally. Hence, being able to read the emotional state of a patient is an important skill. To assess how well students read the VH's emotional state, participants ranked how scared and how friendly they thought the VH was (scale: 1 - 100, 1 = very little, 100 = very much).

4.6.3.3 Medical interview communication

Participants evaluated their medical interview communication with the VH by filling out a survey of twenty five questions. Participants were familiar with the survey as it is commonly used by their instructors to evaluate their interviews with real patients. The survey's continued use by medical instructors indicates it is a reliable evaluation metric for this interpersonal scenario.

Questions from the survey are divided into three categories, information gathering, building rapport with the VH, and procedural aspects of the interview. Sample information gathering questions are "Rate how well you found out all complaints" and "Rate how well you elicited the story and meaning as well as biomedical facts." Sample rapport-building questions are "Did you legitimize the patient's ideas and feelings?" and "Did you use appropriate eye contact?" Sample procedure questions are "Did you use the patient's name appropriately?" and "Rate how well you began with open-ended questions and

moved to closed-ended questions." Questions were rated on four or five point Likert scales (Never, Rarely, Sometimes, Usually, Almost Always or Poor, Fair, Good, Excellent).

4.7 Study Results and Discussion

4.7.1 After-Action Review Impacts Self-Perceptions

Survey results show that after action review changed participant perceptions of their interaction with the VH. Participants indicated they were less friendly ($F_{1,19} = 17.7$, $p < .001$, Pre-AAR: $M = 5.5$, $S_E = .19$, Post-AAR: $M = 4.6$, $S_E = .25$) and less natural ($F_{1,19} = 17.1$, $p < .001$, Pre-AAR: $M = 3.8$, $S_E = .27$, Post-AAR: $M = 2.7$, $S_E = .28$) towards the virtual human than they remembered. On 15 separate measures of rapport-building (e.g., non-verbal behavior, listening, and sensitivity) users indicated their rapport with the VH was worse after the AAR ($F_{1,19} = 18.4$, $p < .001$, Pre-AAR: $M = 2.54$, $S_E = .11$, Post-AAR: $M = 2.18$, $S_E = .11$) . The AAR also changed the way participants interpreted the state of the VH. The VH was perceived as being more scared after AAR ($F_{1,19} = 4.1$, $p < .06$, Pre-AAR: $M = 69$, $S_E = 5.0$, Post-AAR: $M = 76$, $S_E = 2.8$) . No differences were found on ratings of the user's emotional expression and on perceptions of the VH's friendliness.

The effect of AAR on perceptions of information gathering and procedural skills was not as clear. After AAR, participants rated themselves lower on "finding out all the patient's complaints" ($F_{1,19} = 18.3$, $p < .001$, Pre-AAR: $M = 2.4$, $S_E = .15$, Post-AAR: $M = 2.0$, $S_E = .12$) and "beginning with open-ended questions and moving to closed-ended questions" ($F_{1,19} = 4.87$, $p < .04$, Pre-AAR: $M = 1.8$, $S_E = .18$, Post-AAR: $M = 1.5$, $S_E = .17$) . In contrast, differences on "gathering the patient's story as well as biomedical facts" and "using the patient's name appropriately," were small and not statistically significant.

Participant comments support the notion that AAR helped users gain awareness of how they behaved in the VH interview. One participant said, "I like this part [IPSViz] ... actually seeing it. Seeing where my eyes were. Seeing like, having

[the transcript] written out for me - she did this, and then I responded this way. It helps a lot more to know how you did."

Students mentioned specific improper behaviors that IPSViz helped them become aware of.

- **Gaze:** "I didn't realize how much I was kinda looking around. Like, I switched it to her point of view and I was kind of looking all around instead of looking at her."
- **Nervous behavior:** "I think that I learned that kinda shifting in the chair can be easily recognized by the patient as a sign of maybe discomfort or something like that."
- **Response to empathy challenges:** "From looking at the eye contact, I think that was an appropriate level of nonverbal stuff, but I didn't give any verbal empathetic responses on reflection, on looking back at the playback."
- **Information gathering:** "[It was] interesting to see that I just barely hit some of the required or desired topics - and I didn't really stick with anything for that long."

These comments highlight that the visualizations provided by IPSViz were key in making students more self-aware. The 3D rendering of the interaction - including the gaze target, the user model, and the patient's viewpoint - helped students understand how their gaze, posture, and nervous behavior is interpreted by patients. The transcript and topic plot helped students recall the structure and content of the interview, as well as determine what questions they forgot to ask the VH. Finally, video and audio rounded out the feedback provided by the other visualizations by allowing users to watch and hear their inappropriate responses to the VH.

4.7.2 Changing Communication with Real Humans

Students reported they would change behavior in future interviews with real patients based on the AAR with IPSViz.

- **Gaze:** "Definitely, it will make me be more conscious of where I'm looking when I'm talking to a patient."
- **Response to empathy challenges:** "I guess being more mindful about what I'm actually saying. Because when she mentioned breast mass I went directly to, you

know, where is it, you know, tell me about the breast mass, rather than that must be concerning to you. And I think I learned from that."

- **Overall behavior:** "It was good to watch myself. See how I act, and to hear my voice, how I ask questions, the [in]tonation in my voice. And after doing it I can think about it more, you know, things I could have done. And then when I go back and view, I can see ways I could have changed."

The notion that an H-VH experience should impact future interactions with real humans is a surprising result of the study. Given the technological limitations of the system (e.g., speech recognition and understanding), it was not clear that users would see the VH experience as correlating to interactions with real patients. **AAR served as a catalyst to connecting interviews with the VH to interviews with real humans.**

4.7.3 AAR, VH Skin Tone, and Display type

As mentioned previously, two separate studies were run concurrently with the AAR study to evaluate the effect of display type and VH skin tone on the interaction. This section discusses interaction effects between AAR, skin tone and display type.

Display type did not affect the AAR responses, but an intriguing cross-interaction between VH skin tone and AAR was found. As shown in Figure 4-8, the difference due to AAR on "showing interest in the VH" ($F_{1,19} = 6.9$, $p = .02$) was affected by the VH's skin tone. After AAR, students who talked to a dark-skinned VH lowered their score on showing interest to the VH. Students who interacted with a light-skinned VH did not change their score.

The cross interaction between the VH Skin and AAR conditions shows that participants considered the skin tone of the VH when evaluating their behavior with IPSViz. They evaluated themselves differently on this measure depending on whether they interacted with a dark or light-skinned VH. This implies participants (consciously or subconsciously) perceived that the skin color of the dark-skinned VH biased their interaction. Thus, the VH experience served as a bias detector. By interacting with a VH and conducting an AAR, participants detected 1) a bias, and 2) that this bias affects their behavior.

Detecting one's biases is a first step towards changing those biases. Thus, if a VH experience can help a person detect a real-world bias, a VH experience may also be an effective tool for changing that real-world bias. We are currently exploring the interplay between after-action review and many types of biases (e.g., skin tone, ethnicity, religion, weight, age) with the goal of building virtual human experiences for diversity training.

4.7.4 Scope of Perceptions

While these results highlight ways VH experiences can change perceptions of one's communication skills, the full *scope* of these perceptions is not clearly defined. This section conducts a deeper analysis of participant comments, with the goal of expanding the known scope of impacts VH experiences can provide. Defining the scope of these perceptions is crucial to understanding the range of impacts virtual human experiences can have on users.

The section is divided into four parts, each discussing a different group of communication skills that users highlighted during the AAR. All of these communication skills are not just important in this medical training scenario, but for *any social context*. They are:

- Verbal communication
- Nonverbal communication
- Establishing Rapport
- Communicating under stress

In identifying user reflection on these broad communication skills, we show that interacting with a VH and conducting an AAR of that interaction enables evaluating a range of one's communication skills. Furthermore, as these skills are useful in many social contexts, the lessons learned from the VH interaction may apply outside the social scenario the VH interaction simulates.

A common theme in the sections to follow is that users were willing to criticize their communication with a virtual human as if it was communication with a real human. This means users saw virtual humans as equivalent, or as an acceptable proxies for, real humans. The notion that a virtual human is an acceptable proxy for a real human

for evaluating one's communication skills shows the viability of virtual humans for communication skills training.

4.7.4.1 Verbal communication

AARs of H-VH experiences help users evaluate their verbal communication skills. Verbal behavior is crucial to communicating information to others as well as gathering information from others. For example, in the patient-doctor interaction, doctors must ask the patient questions to diagnose her. Likewise, the doctor answers the patient's questions verbally (as opposed to nonverbally, for example, with a nod). Gathering information well enables doctors to diagnose patients faster and more accurately. Likewise, answering patient questions clearly is also important. Patient questions often lead to new topics of the interview that are diagnostically relevant. Also, answering patient question helps patients understand their situation. The relationship between effective verbal communication and diagnosis makes it arguably the most important communication skills for a doctor to learn.

In this study, the AAR helped participants note their difficulty with verbal communication. One participant directly stated, "[I did] so-so on the question answering/asking part." Three specific aspects of verbal communication were highlighted as needing improvement, the phrasing of questions, the thoroughness of the interview, and its organization.

Questions phrasing: Participants noted that the phrasing of their questions affected what information they got from the patient, or how quickly that information was gathered: "You definitely see things as soon as you watch it a second time that I wish I asked it that way."

Thoroughness: Several participants commented on the thoroughness or completeness of the information they gathered. They felt they did not ask all the questions they needed to: "There are some things that I guess I saw that I should have asked or didn't ask."

The topic plot played a role in helping students evaluate thoroughness. The topic plot graphically depicted when important scenario topics (e.g., breast mass, family medical

history) were discussed. If a topic was not discussed, the corresponding row for that topic contained no marking. Thus, students were able to quickly identify questions they forgot to ask, or did not ask enough of, by looking a the topic plot: "The topic graph was a really interesting thing...Interesting to see that I just barely hit some of the required or desired topics, and didn't really stick with anything for that long."

Organization of interaction: Collecting information in a logical, orderly fashion helps with organizing the information gathered. Furthermore, demonstrating organization to others is important in interpersonal interactions. Organization shows competence and preparedness to others. In addition, organization helps the conversation partner understand where the conversation is going.

Several participants noted their interviews were organized poorly. One participant stated, "I noticed I was jumping around - asking questions about the breast lump, and then I would try go to past medical history, and then I would, like, try to go back."

Again, the topic plot played an important role in shaping participant perceptions of their organization. For some, the topic plot introduced the idea of interview organization: "There were things after looking at the review tool. especially like the little plot with the flow of conversation. I guess I never thought about the flow of conversation through the interview." For those who already understood the concept of interview organization, the topic plot gave them a better understanding of what an organized interview looks like: "I think one of the things I did like is the [topic plot's] color coding of what particular parts of the history were asked at different times. I guess you would sometimes want all the colors to be in a similar place, if you were doing a pretty systematic approach." One participant compared what she considered the ideal organization of an interview to what he observed on the topic plot of his interview: "Normally you're going to do your chief complaint, history, blah blah. But showing things being all around [on the topic plot], pain here, symptoms here, pain back again. That gets you kind of seeing how the interview

flows or did not flow by that usage." Thus, the topic plot helped this participant see that she was not following the ideal organization.

4.7.4.2 Nonverbal communication

The AAR impacted participant perceptions of three kinds of nonverbal behavior, gaze behavior, body postures, and paralanguage. These behaviors have been shown to communicate a variety of attitudes that are important in all social interactions, including attention, engagement, interest, nervousness, dominance, confidence, and liking (or the lack thereof) [8, 124].

Gaze behavior: Participants said gaze behavior was poorer than they remembered: "It's kind of surprising seeing where my eyes actually were . . . I was probably a little lax about it . . . I was looking at the patient but I wasn't maintaining eye contact. I was looking just kind of in that general area . . . I'll be more conscious about my eye - where my eyes are focused." Participants not only considered where they were looking, but also the frequency with which gaze changed. One participant realized the lack of changes in her gaze reflected a lack of engagement with the virtual human: "This, oh this is way more high tech. As far as, like, looking, watching where you're looking. At the last one, I kind of just tuned out. I wasn't engaging her so much as just staring straight ahead. Not like I'm talking to you. I look you in eye and you know . . ."

Participants evaluated gaze behavior by observing both the gaze target and the avatar's head orientation: "I didn't realize how much I was kind of looking around. Like, I switched it to her point of view and I was kind of looking all around instead of looking at her . . . [This] will make me be more conscious of where I'm looking when I'm talking to a patient. Because I thought I had fairly good eye contact as far as paying attention to her, but the bullseye was going all around. I saw that."

The comments also show that users can interpret AAR visualizations incorrectly. In several of the quotes above, end-users misinterpreted the meaning of the gaze target. Although the gaze target demonstrated head gaze, several end-users thought the gaze

target demonstrated eye gaze. Some even said they would change eye gaze behavior with real humans as a result of this experience.

Misinterpretation of the AAR visualizations presents a challenge. If end-users misunderstand the visualization, they may be encouraged to change behavior that does not need to be changed (or vice versa). Efforts should be made to disambiguate the meaning of visualizations so that end-users do not change behavior they do not need to. Nonetheless, it is still encouraging that the gaze target impacted end-user perceptions. This shows that AARs have *the capacity* to change perceptions of behavior. Capitalizing on this capacity to change perceptions *accurately* is an important step in developing effective AAR tools for H-VH experiences.

Paralanguage: Paralanguage are vocal expressions that are not speech. They include tone of voice, accents, pronunciation, and pauses in speech. Paralanguage can also communicate subtle attitudes and feelings to others, such as excitement, disdain, sympathy, and nervousness [78]. Thus, evaluating one's paralanguage is important for learning how one communicates with others.

During AAR, the following forms of paralanguage were identified as needing improvement.

- **Accents and pronunciation:** "I think I was getting impatient because I kept having to repeat and rephrase [questions] . . . I have a feeling some of that was my accent and some of that was poor articulation."
- **Tone of voice:** "It was good to watch myself, see how I act, and to hear my voice, how I ask questions, the tonation [sic] in my voice. And after doing it I can think about it more, you know, things I could have done. And then when I go back and review, I can see ways I could have changed."
- **Pauses in speech:** "I don't think I realized that pauses would have been useful until now. I just kind of forged through...I didn't use some of the skills I use in the clinic normally - pausing and waiting for the patient to answer."

Paralanguage played a large role in communicating discomfort and nervousness to the VH. For example, long pauses (or quiet) communicated discomfort and lack of friendliness:

"I learned] sometimes when you're thinking and, you know, being quiet for a little bit, it seems like you're uncomfortable or cold or unfriendly." Furthermore, tone of voice communicated nervousness: "I realized that I was not comfortable at all and that kind of came across, and I sounded nervous, and usually I don't sound like that at all."

Body posture and motion: Participants indicated their body posture and motion was detached from the interaction: "When I watch myself back I realize like I didn't participate physically at all." The body lean plot played a role in shaping this perception: "In the chart of my lean, you can kinda see how that reflects . . . I guess I kind of always lean forward at the same angle, but when you just kind of do this [leans forward], it picked up on that and I thought that was kind of interesting." Another participant found that his nonverbal behavior communicated nervousness: "I think that I learned that kind of shifting in the chair can be easily recognized by the patient as a sign of maybe discomfort or something like that."

Reviewing the user's avatar from the patient's perspective also helped participants evaluate their body posture and motion: "It's interesting seeing where your head starts or your body position–that you don't realize while you're interacting with the patient. Like I think that might be one of my favorite features of [IPSViz], getting to see how I was positioned, where my head was turned. And I like being able to see it from the patient's perspective, too, what it looks like, because that's a neat tool and something you don't think about." This comment highlights the importance of reviewing an experience from the perspective of a conversation partner. It enables seeing how nonverbal behavior might be interpreted by a conversation partner.

4.7.4.3 Establishing a relationship

AARs of VH experiences help users evaluate their relationship-building skills. In many social contexts, establishing a relationship with a social partner is crucial to meeting one's goals. For example, in the patient-doctor interaction, doctors that establish relationships with patients 1) are sued for malpractice less and 2) have patients with faster

healing times and better outcomes. Thus, relationship-building is an important skill for a medical student to learn. In this study, the AAR helped student note their poor greetings, lack of empathy, and poor listening skills, each of which hurt their ability to establish a relationship with the patient.

Greeting and introduction: A key aspect of relationship-building is the greeting and introduction. It is considered rude to not introduce oneself or greet another person. During the AAR, some participants observed that they did not greet or introduce themselves to the patient: "I kind of forgot to do the whole process of 'hello,' and then, 'this is what we're going to do next.' "

Empathy: Colloquially, empathy is the ability to put oneself in another's shoes. Empathy builds relationships by showing a conversation partner that you understand their perspective. Understanding another person's perspective is crucial for reaching common ground, gaining trust, and making group decisions [42, 90, 101]. Many participants said they were not as empathic as they thought: "There was definitely a lack of empathy and actually relating to the patient." One participant said he missed opportunities to be empathic: "I feel like I learned, um, like how to, you know, if the person brings up it could possibly be cancer, then I should probably address that. I think I was too focused on trying to get what kind of symptoms she has." The student not only recognized that he wasn't empathic, but that his strong desire to diagnose the patient prevented him from recognizing the patient's need for empathy.

Listening skills: Listening is an important skill in a social interaction. Maybe more important than listening is *demonstrating* that one is listening. By demonstrating listening, social actors show they care about what others are saying, and that what others are saying is important. Users noted they missed important moments in the interaction, including when the patient said she was nervous (a good opportunity to be empathic), and when the patient sneezed: "[I learned] maybe to pick up more subtle clues of patients, like when they say that they're nervous from the beginning, instead of just trying to plow right

through while they're there - acknowledging that and even little things like picking up on when they sneeze."

4.7.4.4 Reactions under stress

Stressful situations may occur in a social context. Learning how to be calm and focused in such situations can be important to meeting one's goals.

During the AAR, participants noted their reactions to the virtual human while under stress: "I have a lot to improve on. I'm a first-year PA student so I've had limited patient experience. Especially going into something where you don't really have any idea of what you're going to get into. So it was a little bit of a struggle. And then as soon as I got flustered, of course, it increased just as my own awkwardness already at doing [a patient] history as well. That's good to see in this study."

A common theme in participant discussions about stress was that the stress stemmed from interacting with the virtual human: "Also it shows that, I guess under a little bit of stress . . . you're in there with all the cameras and what not, I guess you can see how things don't always flow talking with the patient."

In VH experiences, stress level is particularly high because of the technological limitations of today's VHs. With our VH, the VH does not always respond to questions correctly or display emotions properly. Furthermore, the interaction requires encumbrances such as tracking equipment. These technological limitations frustrate and stress end-users. While the causes of this stress are different from the causes of stress in a similar real-world scenario, multiple participants said it was beneficial to review how they react under stress with the VH. This indicates participants saw a relationship between their behavior under stress with a VH, and their behavior under stress with a real human.

4.7.5 Summary

Through AARs, H-VH experiences drive self-awareness and self-directed change of interpersonal skills. In the study, participants conducted an AAR of their own interaction with a VH. The AAR either changed participant perceptions of their interpersonal skills,

or at least drove participants to reflect on them. In some cases, users spontaneously professed a desire to change behavior with real humans as a result of the AAR.

The scope of real-world interpersonal skills observable in an H-VH experience are broad enough for use in evaluating real-world interpersonal skills. Participants observed and reflected on their verbal and nonverbal behavior, their ability to build a relationship with a communication partner, and how well they communicated under stress. These skills are important in H-H interactions. Thus, not only can reviewing an interaction with a virtual human provide insight into one's interactions with real humans, but that insight pertains directly to skills that matter in H-H experiences.

4.8 Impact of Interaction Realism on AAR Effectiveness

Although AARs of H-VH experiences can provide insights into skills that matter in H-H experiences, the study unexpectedly uncovered a factor that can attenuate these insights: VH interaction realism. The less realistic an interaction with a virtual human (in comparison to an interaction with a real human), the less an AAR of that interaction will impact self-perceptions of interpersonal skills. Self-perceptions are not impacted as strongly because participants see an interaction with an unrealistic virtual human as unrepresentative of their interactions with real humans. For these participants, there is little connection between their demonstrated interpersonal skills with a virtual human and their actual interpersonal skills with a real human.

Thus, to elicit the *most* impact from a human-virtual human experience, one should 1) add AARs to the experience, and 2) strive for high interaction realism. Note that striving for high interaction realism is not the same as requiring high interaction realism. The study results indicate the realism of the interaction does not need to match that of a real human interaction. **Lower levels of interaction realism still elicited self-reflection, changes in self-perceptions, and self-directed change from the AAR.**

The remainder of this section mirrors the process by which the relationship between VH interaction realism and the effectiveness of the AAR was identified. In the first part of the section, we review multiple comments from participants that suggest the relationship exists. The comments highlight 1) the VH's low interaction realism (as compared to a real human), and 2) how low interaction realism, and not other forms of realism, affected views of the usefulness of the AAR. These comments led to the following hypothesis: The less realistic an interaction with a virtual human (in comparison to an interaction with a real human), the less an AAR of that interaction will impact self-perceptions of interpersonal skills. In the second part of the section, we operationalize VH interaction realism and changes in self-perceptions due to the AAR to statistically confirm the hypothesis.

4.8.1 Realism vs. AAR Effectiveness: Participant Comments

After the AAR, participants said the virtual human did not meet the standard of interaction realism set by real humans. Two specific aspects of interaction realism were cited by participants, the virtual human's difficulty understanding speech and the virtual human's lack of nonverbal expression. 18 out of 27 participants mentioned the virtual human's difficulty understanding their speech. For example, one participant said, "It didn't feel like a natural conversation flow, and I had trouble with her like, I had trouble, she had trouble understanding me I think sometimes." In addition, 4 out of 27 participants mentioned the virtual human did not communicate nonverbally, making it difficult to read her emotions and attitudes as they would in an interaction with a real person. One participant said the lack of facial expressions made it difficult to be empathetic with the VH: "I think you guys do a good job with making the patient fairly lifelike, but you still can't get a lot of facial expression so it's a little bit harder to read emotions. So I think - and I don't have video tape [of myself] to back this up - but I think with real patients its a little bit easier to be empathetic." This participant believes that she would have been more empathetic with a real human because a real human would express her fears both verbally and nonverbally. Other comments about nonverbal

behavior were more subtle. For example, one participant said, "I couldn't tell if she was really answering my question, like she understood my question." The participant's comment not only reflects the virtual human's difficulty answering questions, but also the lack of nonverbal behavior, such as nodding, that is expected in real human interactions.

Since participants saw the interaction with the virtual human as unrepresentative of real-world interpersonal skills, participants dismissed some of the feedback from the AAR as irrelevant for interactions with real humans. 26 out of 27 participants saw themselves interact poorly with the virtual human, but dismissed that feedback partially or entirely due to the VH's lack of human realism. For example, one participant saw the virtual human as more like a program than a human: "I treated it as a program instead of trying to get to the diagnosis. [I did] not really treat the patient as a person as I usually would treat if a real person comes to the clinic instead of a simulated program." To this participant, his behavior with "a program" is not a reflection of his behavior with real people. One participant indicated the VH's lack of realism kept him from using a specific the interpersonal skill of empathy: "When I went in there I didn't think so much about using empathy, because this is a computer patient. I noticed that when I looked at it again, that's very different than how I am with a real patient." Another participant dismissed his mistakes with the virtual human because of the VH's difficulty answering his questions: "I don't guess I would do anything any different [in a future interaction with a patient]. Because that's not really how I would go into a room. It's difficult when they don't understand what you're saying. You have to ask them [questions in] certain ways." This participant felt he should not change his behavior in future interactions with real patients as a result of the AAR, because the AAR provided him with feedback based on an unrealistic situation.

Despite the apparent effect of VH interaction realism, all participants self-reflected on their interaction with the virtual human and many found that the interaction with the virtual human provided *some* indications of their interpersonal skills with real humans.

Of the 26 users who highlighted virtual human realism as a problem, 11 still expressed personal responsibility for their mistakes with the VH patient. One participant said, "I didn't catch the sneeze and that's something. A lot of times in real life I would catch it but I might - I mean, I can't - I don't know whether its the simulation or whether its just hard to write it off as that. It's easy for me to write off all this as 'oh it's just a simulation' but I've never been videotaped talking to a real person either." Another expressed his frustration with the VH, but ultimately felt he might make the same mistakes with a real human patient: "Because, I couldn't understand her, she couldn't understand me. And then by the time it started making sense, I was frustrated, so I don't think I responded the way I would have ... But at the same time, I don't know how much better my responses would have been with a normal patient. I think that has a lot to do with being a first-year." In spite of the VH's deficiencies, these participants verbalized a connection between their behavior with the VH and their behavior with real humans, such that they could not completely dismiss their observations in the AAR as invalid.

The focus of participant comments on VH interaction realism supports the argument that interaction realism is more important than other forms of realism in H-VH experiences [33, 189]. Other forms of realism, such as visual realism and scenario realism, were not mentioned by participants. This indicates other forms of realism were a minimal factor (in comparison to interaction realism), or not a factor at all, in determining the effectiveness of the AAR.

4.8.2 Realism vs. AAR Effectiveness: Statistical Analysis

Based on the anecdotal evidence in the previous section, it appears that increasing VH interaction realism will improve the effectiveness of the AAR (and vice-versa). This sections aims to confirm this relationship between VH interaction realism and AAR effectiveness statistically. The following hypotheses about the relationship between VH interaction realism and AAR effectiveness are proposed:

- **H1:** VH interaction realism is positively correlated, and thus predictive, of AAR effectiveness.
- **H2:** VH interaction realism is a factor in determining AAR effectiveness.

These hypotheses are confirmed in the following subsections, bolstering the argument that VH interaction realism has a larger influence over the success of an H-VH interaction than other forms of realism, as well as motivating the need for new approaches to creating virtual humans that interact more like real humans than today's state-of-the-art.

4.8.2.1 Operationalization of VH interaction realism and AAR effectiveness

To confirm these hypotheses, the subjective concepts of VH interaction realism and AAR effectiveness must first be operationalized. **VH Interaction Realism:** As participants singled out the virtual human's speech understanding as the primary factor affecting the interaction, VH interaction realism is operationalized as R, the speech understanding success rate. The speech understanding success rate is the percentage of appropriate responses spoken by the virtual human in response to participant questions. VH speech accuracy was not a controlled study variable, but it varied enough from participant to participant (M=74%, SD=8.4%, min=56%, max=86%) to allow a meaningful analysis of its relationship with participant self-ratings. The variation in VH speech understanding was a result of several factors, including differences in participant speech patterns, accents which the speech recognition software had difficulty understanding, and the extent to which each participant asked the VH questions for which the VH had an answer.

As it is difficult to algorithmically determine if a VH responded properly to a person's speech, R was calculated by external observers. Observers first transcribed each H-VH interaction. Then, for each participant utterance in the interaction, the virtual human's corresponding responses were categorized as logical or illogical for the utterance. For example, if the participant said, "Do you have any allergies?" and the patient responded, "My mother had breast cancer," this would be categorized as illogical because the

VH's response does not follow from the participant's question. If the VH responds, "I don't have any allergies," then the response would be categorized as logical. The speech accuracy success rate for each interaction is then computed as the percentage of logical VH responses among total VH responses.

AAR Effectiveness: As an AAR can be viewed as effective if it changes self-perceptions, AAR effectiveness was operationalized as $\Delta_{Pre-AAR}^{Post-AAR}$, the change in pre-AAR self-ratings of performance with the VH as a result of the AAR. Self-ratings on a variety of interpersonal skills were taken from participants by administering a survey before and after (Section 4.6.3) the AAR. If self-ratings change, the implication would be that the AAR was effective as eliciting self-reflection on an H-VH interaction, and changes in self-perceptions of interpersonal skills.

Using R and $\Delta_{Pre-AAR}^{Post-AAR}$, two types of statistical tests were conducted to confirm hypotheses H1 and H2, correlations (Pearson's) and difference tests (ANOVA), respectively. The correlation tests show that VH interaction realism predicts AAR effectiveness on providing feedback on rapport-building. The difference tests show VH interaction realism affects the feedback from an AAR on rapport-building

4.8.2.2 H1: VH interaction realism is positively correlated with AAR effectiveness

Identifying a positive correlation between VH interaction realism and AAR effectiveness has several benefits.

1. It would indicate there is *a* relationship between these variables, bolstering the argument for virtual humans that interact more like real humans.

2. It would enable prediction of AAR effectiveness based on the realism of the VH interaction. If VH interaction realism is low, it may not even be worthwhile to conduct an AAR of the interaction. (Note: Even if VH interaction realism is low, it may still be worth conducting the H-VH interaction itself. The benefits of the H-VH experience do not come from the AAR alone. There are benefits in conducting the experience itself [86].)

3. It would enable measuring VH interaction realism indirectly. Realism is an abstract, subjective concept which is more difficult to measure than AAR effectivness. If AAR

effectiveness is correlated with VH interaction realism, then one could use AAR effectiveness to measure realism.

Two correlation tests show the effectiveness of reviewing rapport-building skills are positively correlated with VH interaction realism (as measured by the speech understanding success rate). Thus, if VH interaction realism is too low, it will likely be difficult for users to get feedback from an interaction with a virtual human on the crucial skill of rapport-building. In the medical context of this study, feedback on rapport-building skills were particularly crucial to the study participants because better rapport with patients typically leads to better health outcomes for patients and less malpractice lawsuits. Good rapport likely also provides similar benefits in other interpersonal scenarios. Thus, improving VH interaction realism may enable users of VH experiences to get effective feedback on rapport-building skills.

The first correlation test was made between the VH speech success rate R and the change in self-ratings due to AAR $\Delta_{Pre-AAR}^{Post-AAR}$. A negative correlation was expected, meaning participants who encounter a more realistic VH (a VH that speaks more like a real human) would rate themselves poorer on rapport-building post-AAR than those who saw a less realistic VH. A negative correlation was found on a measure of rapport building, whether the participant "expressed support and partnership" to the VH (Two-tailed, Pearson's r = -0.382, p = 0.049). In other words, as VH interaction realism increased, participant ratings of themselves on support and partnership trended lower post-AAR. This means participants who saw more realistic VHs were more willing to trust the feedback provide by the AAR on their use of rapport-building.

The second correlation test was made between the VH speech accuracy rate and the abs($\Delta_{Pre-AAR}^{Post-AAR}$), the *magnitude* of the difference between participant self-ratings before and after the AAR. A positive correlation was expected, meaning participants who encounter a more realistic VH would change their self-ratings on rapport-building more than those who saw a less realistic VH. A positive correlation was found on two measures

of rapport-building, whether the participant "demonstrated interest and respect" (r = 0.468, p=0.007) and a rating of the participant's use of empathy (r = 0.398, p =0.374). In other words, as VH interaction realism increased, participant ratings of themselves on demonstrating interest and respect and using empathy changed more post-AAR. Again, this shows participants who saw more realistic VHs were more willing to trust the feedback provide by the AAR on their use of rapport-building.

Although unrelated to rapport-building, a correlation was found that indicates frustration with the virtual human was related to the virtual human's interaction realism. A negative correlation was found between VH speech accuracy and the extent to which participants though they were emotional when interacting with the VH (r = -0.362, p = 0.032). In other words, as the interaction realism of the VH increased, participants tended to not change self-ratings of their emotional expression. This is likely because participants who talked to less realistic VH's were more likely to be frustrated by the VH, and therefore more likely to express that frustration emotionally when talking to the VH.

In summary, it appears VH interaction realism is correlated specifically to how one evaluates their rapport-building with a virtual human. The less realistic the interaction with VH, the less willing users will be to change evaluations of their rapport-building with a VH. If we aim to improve rapport-building skills using H-VH experiences combined with after-action reviews, then the interaction realism of the virtual human will provide a good predictor of the success of the AAR.

4.8.2.3 H2: VH interaction realism is a factor in determining AAR effectiveness

Building on the correlative relationship identified previously, this section shows that the effectiveness of reviewing rapport-building skills demonstrated in a H-VH experience is *affected* by the level of VH interaction realism (as measured by the speech understanding success rate). This means users of H-VH experiences will not get valuable feedback on rapport-building skills from interactions with virtual humans. This further emphasizes the need to improve VH interaction realism.

To determine if low VH interaction realism is a factor in determining AAR effectiveness, participants were grouped into one of four increasing levels of speech accuracy and a one-way ANOVA on $abs(\Delta_{Pre-AAR}^{Post-AAR})$ was computed. If VH interaction realism is factor in determining AAR effectiveness, then participants with lower speech accuracies will have significantly smaller changes in self-ratings as a result of the AAR.

The groups were determined by binning speech success rates by standard deviations from the mean speech success rate. Let r_i be the speech success rate for participant i's interaction with the VH, m be the mean speech success rate, and σ be the standard deviation of the speech success rate. Then r_i is binned into one of the following four speech success rate groups, where n is the number of participants in each group:

- Group 1 (n=5): $a_i < m - \sigma$
- Group 2 (n=9): $m - \sigma <= a_i < m$
- Group 3 (n=9): $m <= a_i <= m + \sigma$
- Group 4 (n=4): $a_i > m + \sigma$.

The ANOVA shows VH interaction realism affected participant ratings of themselves on rapport (Figure 4-9). The speech success rate group had a significant effect for several measures of rapport, including a summary measure composed from 15 measures of rapport. Furthermore, changes in self-ratings were smaller for those who interacted with a less realistic VH and vice-versa. Thus, the more realistic the interaction presented by the VH, the more participants changed their self-ratings of rapport-building.

Although VH interaction realism affected AAR effectiveness, even participants on who interacted with a VH who demonstrated low interaction realism still changed their ratings due to AAR (Section 4.7.1). This indicates all participants attributed *some* validity or trust to the after-action review. One question this work does not answer is whether there is some minimum VH speech accuracy below which users completely dismiss the AAR as invalid. Clearly, low interaction realism attenuates the impact of the AAR of an H-VH experience. The extent of that attenuation effect is unanswered here.

In summary, VH interaction realism affects how one evaluates their rapport-building with a virtual human. The less realistic the interaction with VH, the less users will change evaluations of their rapport-building with a VH as a result of an AAR. If we aim to improve rapport-building skills using H-VH experiences combined with after-action reviews, then the interaction realism of the virtual human should be measured to make sure users will take the feedback from the AAR seriously.

4.9 Conclusions and Future Work

To what extent (if at all) can we change or affect people using an interaction with a virtual human? This work argues that we *can* change people by combining H-VH interactions with AAR. Guidelines for AAR for H-VH interactions and IPSViz, a tool that generates visualizations for AAR, were presented. In addition, a study demonstrated that combining H-VH interaction with AAR, at least in the short term, changes the way a user sees himself (self-identity and self-awareness). More importantly, user comments indicate this change may extend into future interactions with real humans.

One issue that emerged from this study is that the impact of the AAR can be attenuated by low VH behavioral realism. Low realism led participants to believe that the behavior they observed during the AAR does not accurately reflect their behavior with real humans. Other kinds of realism, such as scenario or visual realism, did not appear to have an effect on participant perceptions post-AAR. Future studies must explore the effect of VH realism closer to determine at what threshold VH realism completely negates the impact of the AAR.

The next step is to learn if using IPSViz results in measurable improvement in H-VH and eventually H-H communication skills. This will be evaluated through repeated exposures to VHs followed by AAR with IPSViz. This experience-feedback loop will enable conclusively evaluating if student comments of AAR changing their perspective is a realizable goal.

To support VH scenario designers, IPSViz is also being modified to review groups of interactions with virtual humans. IPSViz^n characterizes H-VH interactions in the aggregate. Instead of visualizing a single interaction at-a-time, IPSViz^n visualizes n interactions. For example, students can generate a portfolio of H-VH interactions for student and educator review of progression of communication skills; educators can efficiently evaluate class performance on an interaction by reviewing class-wide visualizations; researchers can review how manipulating properties of the virtual human affects users.

IPSViz represents an important progression in the application of VEs. Traditionally, the impact of a virtual experience has been viewed as a product of the experience itself - the VE is what impacts the user and makes the experience valuable. IPSViz extends this to include review and self-reflection. By enabling review and self-reflection, the VH experience continues to impact the student beyond the conclusion of the experience itself. This impact is different than that provided by the experience itself. With AAR, users have an opportunity for self-reflection, insight, and self-directed change for real-world social interactions.

Figure 4-1. Screenshot of IPSViz.

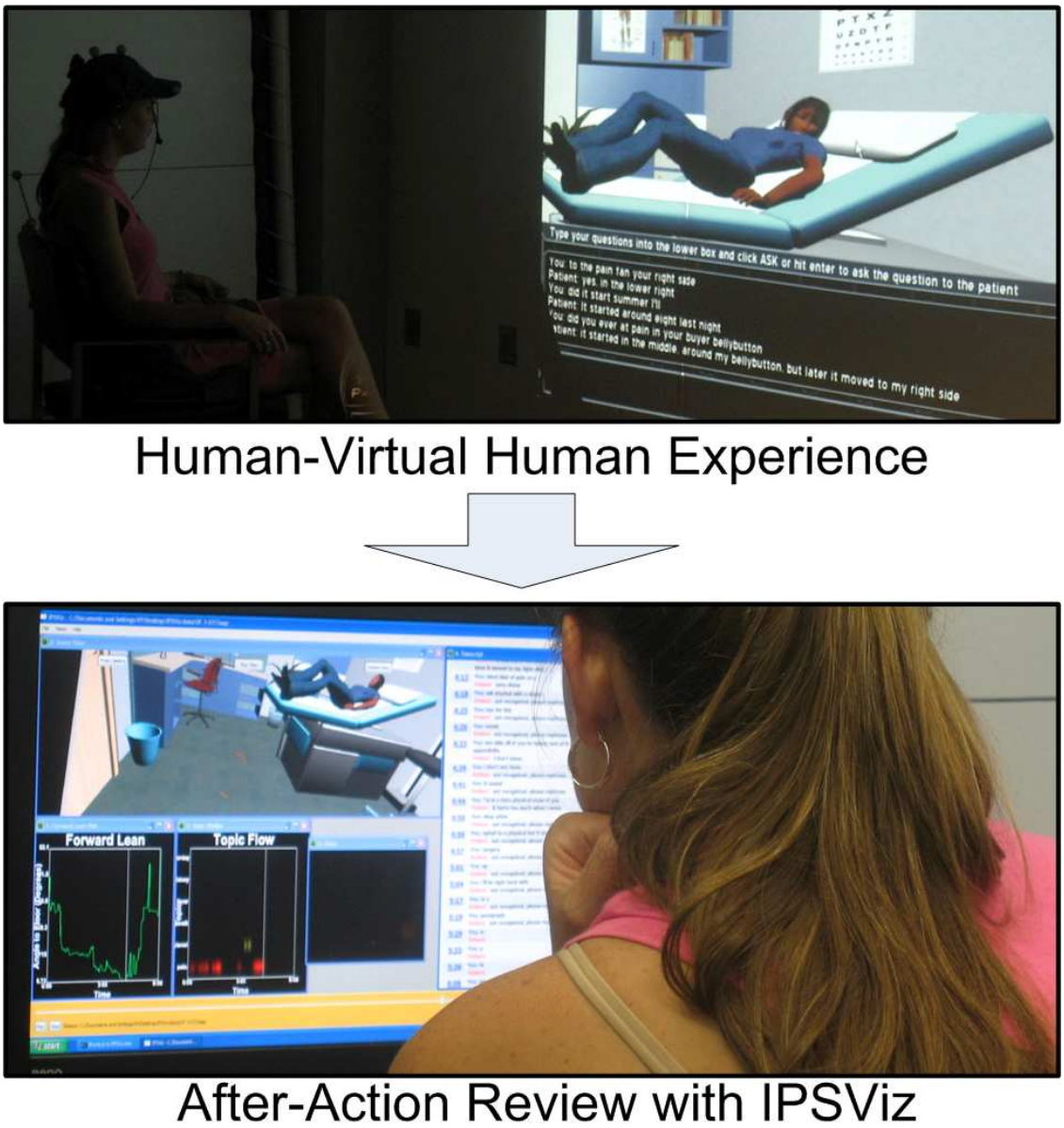


Figure 4-2. User interacts with a virtual human, then reviews interaction with IPSViz.

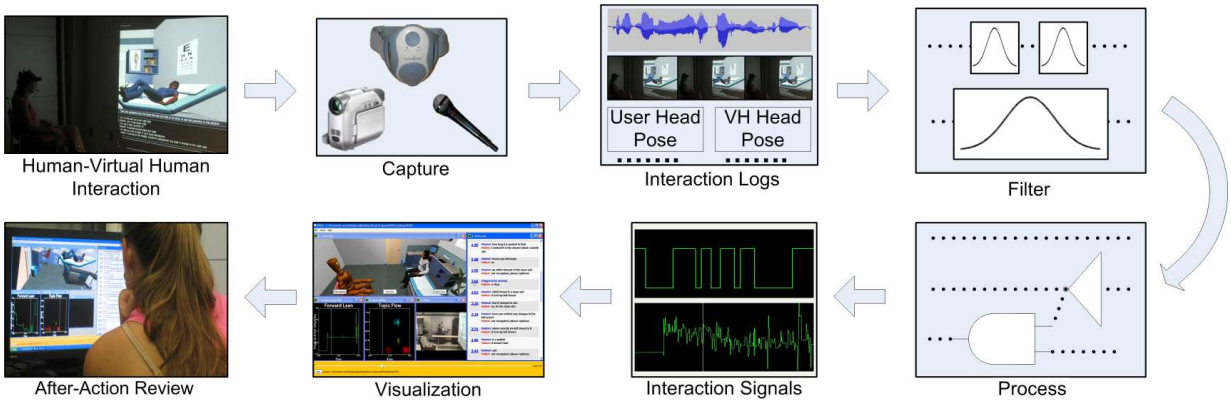


Figure 4-3. H-VH interaction is captured, filtered, processed, and visualized for review, evaluation and feedback.

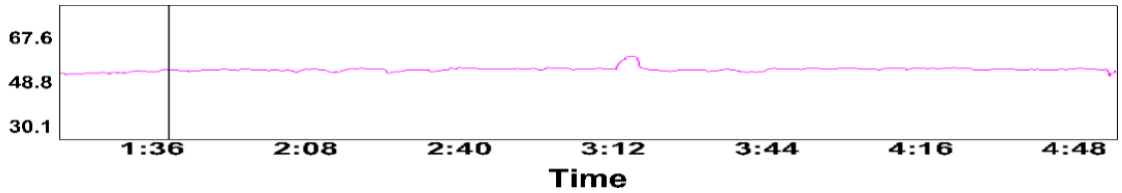


Figure 4-4. Example student body lean throughout an H-VH interaction

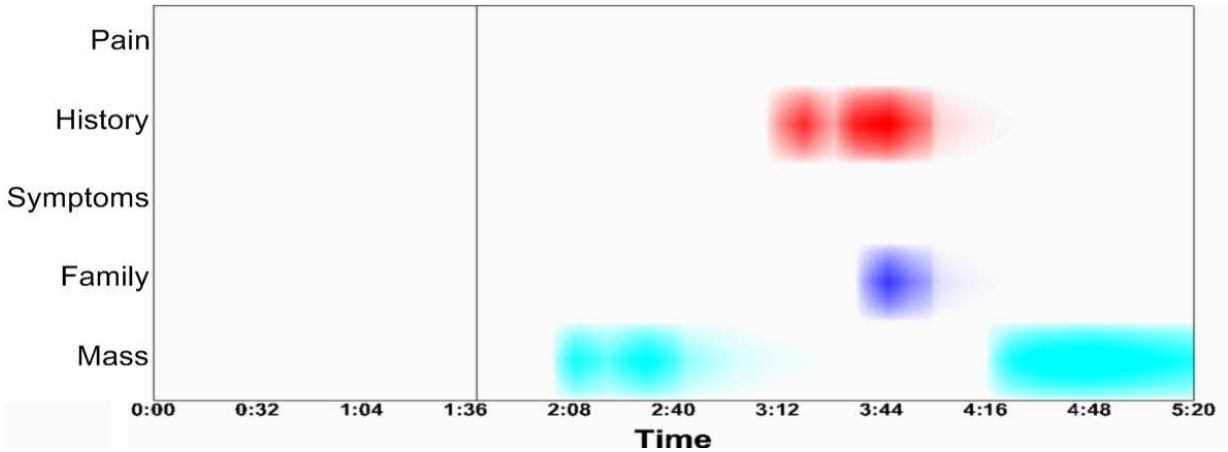
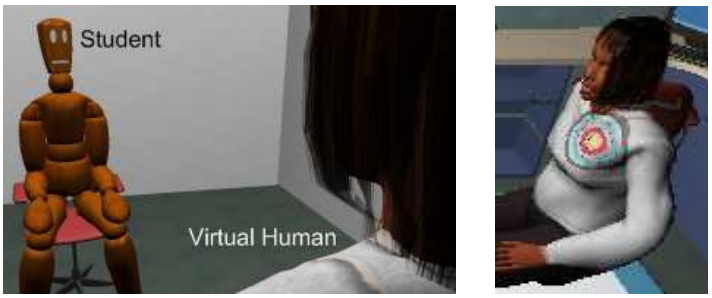


Figure 4-5. Visualization of five topic signals showing when student and VH discussed important topics.



A B

Figure 4-6. User (wooden posing doll), VH and virtual environment rendered in 3D. A) Review from perspective of VH. B) Interaction augmented with gaze information.

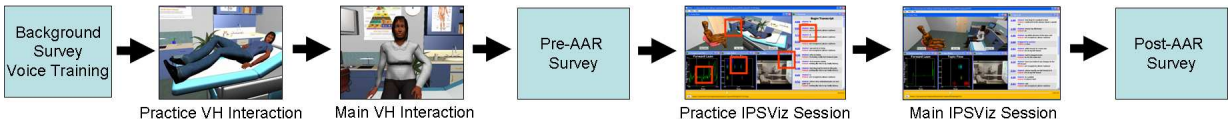


Figure 4-7. Study Procedure

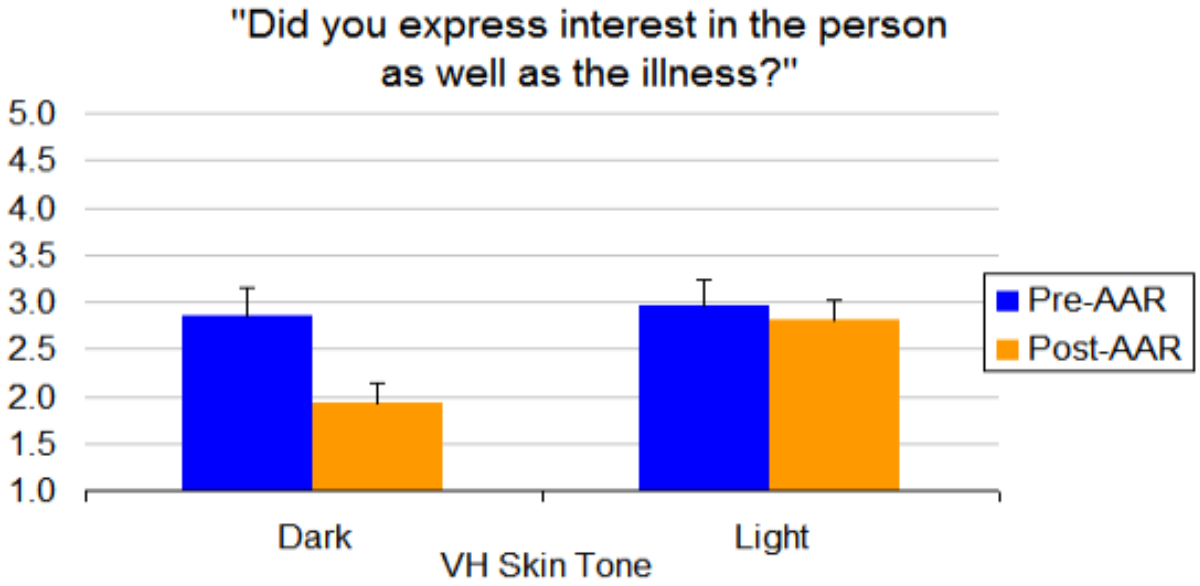


Figure 4-8. Interaction of VH skin tone and after-action review on showing interest in the VH.

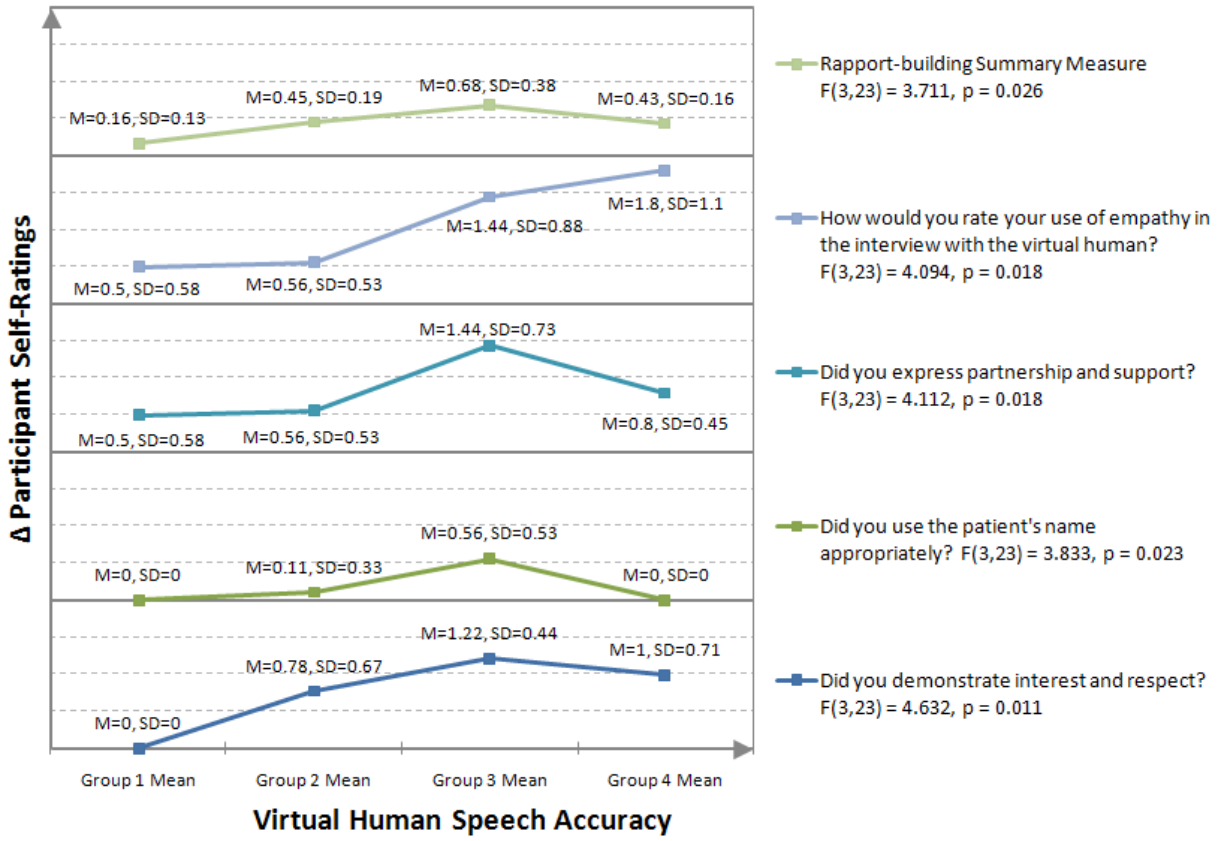


Figure 4-9. VH interaction realism versus change in self-rated rapport-building scores

CHAPTER 5
VSP: VIRTUAL EXPERIENCES FOR SOCIAL PERSPECTIVE-TAKING

This chapter describes the VSP system, an immersive after-action review (AAR) system for interpersonal interactions with a virtual human (VH). VSP leverages virtual reality techniques to enable reliving an interaction with a VH from the VH's perspective. A pilot study reviews the impact of VSP on users of VH experiences. This work will appear at the IEEE conference on Virtual Reality 2009 [146].

Personal contributions: I developed the VSP system in its entirety, designed and conducted the user study, and analyzed the data.

Collaborators: The mixed-reality human used as part of the user study was developed by Aaron Kotranza. Kotranza also assisted with designing and conducting the study, and played a large role in the data analysis. D. Scott Lind provided access to study participants (medical students) as well as expertise on interpersonal communication in the VH scenario (the medical interview).

Relevance to thesis: This chapter describes VSP, the second of three AAR systems for H-VH experiences that were developed as part of this dissertation. The chapter demonstrates the use of first-person perspective for after-action review of an H-VH experience. Furthermore, it shows reviewing one's interaction with a VH from the VH's perspective elicits perspective-taking, as well as self-reflection, changes in self-perceptions, and self-directed change in domain-specific interpersonal skills.

5.1 Introduction

The power of a virtual experience lies in replacing the user's senses with computer-generated sensations to transport the user to another world. We propose a new type of virtual experience that transports the user into a specific experience lived by another person, by 1) replacing the user's sensory inputs with sensory information previously experienced by another person, and 2) prompting the user to reenact this other person's behavior. This

effectively allows the user to live through, reflect on, and learn from the experience of another person.

5.1.1 Virtual Social Perspective-Taking

Social perspective-taking is the process of reflecting on what it is like to be another person. Colloquially, social perspective-taking is often described as "seeing through another's eyes" or "walking in another's shoes." This process helps people understand each other and communicate effectively [42, 90, 101].

We propose virtual social perspective-taking (VSP). Virtual experiences are uniquely capable of facilitating social perspective-taking because they can render an experience from the perspective of another person. Rendering from the perspective of another person enables reliving the experience of another person: seeing what another saw, hearing what another heard, touching what another touched, saying what another said, moving as another moved, and - through narrative and drama - feeling the emotions another felt.

Reliving the experience of another person has the goal of helping users reflect on, understand, and learn from other people's experiences. This goal is realized by using VSP to transport medical students into the potentially unfamiliar experience of being a patient. The patient is a 34-year-old Caucasian woman who is afraid she has breast cancer and is undergoing a physical breast exam. Reliving this patient's unfamiliar experience - particularly unfamiliar for male medical students - improves student understanding of the patient's perspective and student behavior in future patient interactions.

5.1.2 Driving Application: Medical Interview Training

Our exploration of VSP is driven by a real-world application where social perspective-taking is critical to success, the medical interview. The medical interview is a social interaction between doctor and patient where doctors must understand the patient's perspective in order to address the patient's needs, concerns, and fears. In the medical education literature, this is known as expressing empathy.

We developed a VSP system that allows a medical student to relive the experience of one of his virtual patients, Amanda Jones. Amanda arrives at the doctor's office with breast pain and is afraid her pain means she has breast cancer. This breast cancer scenario provides empathetic moments where the doctor should express an understanding of the patient's fears and concerns. Experiencing the medical interview of a fearful patient through VSP is targeted to aid the student in reflecting on and improving his use of empathy with human patients. Amanda is a mixed reality human (MRH) patient [99], capable of interacting with the student through speech and touch, thus enabling the student to interview and examine Amanda much like a real-world doctor-patient interaction. Cameras and logs record the interaction from the patient's perspective for VSP (Figure 5-2).

Immediately following the doctor-patient interaction, the medical student participates in VSP by reliving the recorded interaction as Amanda (Figure 5-1). The student wears an HMD, is seated on a physical exam bed where Amanda sat, and his body is replaced with Amanda's avatar. Three basic principles are applied to allow the user to relive the interaction as Amanda.

1. The student's senses are replaced with recordings of Amanda's senses. The student sees and hears himself asking questions and conducting a physical breast exam from Amanda's perspective.

2. The student is reminded that he is Amanda. When looking at his body or in a virtual mirror, the student sees Amanda's avatar.

3. The student reenacts Amanda's behavior. To prompt the student to reenact Amanda's behavior, the student's avatar is updated with Amanda's movements, and the text of Amanda's speech is displayed in the HMD.

5.1.3 Evaluation

A pilot study (n = 16) evaluated if VSP leads to 1) reflection on another person's perspective and 2) self-directed changes in empathic behavior. Participants interacted with a MRH patient (MRH1), then had a VSP experience of MRH1, and finished with a second MRH patient interaction (MRH2). The VSP experience motivated participants to reflect

on their behavior in MRH1 from the perspective of the patient. This reflection improved participant empathy in MRH2, indicating that VSP elicits social perspective-taking and self-directed improvements in behavior.

5.2 Previous Work

5.2.1 Social VEs Can Benefit from VSP

There has been recent growth in using virtual experiences to simulate social interactions such as conflict resolution [79, 176], cultural competency [10, 40], medical diagnosis [89], and dealing with the mentally ill [59]. These virtual experiences simulate social interactions in which the user can benefit greatly from engaging in social perspective-taking. social perspective-taking has been shown to aid in resolving conflicts [42], promoting cooperation [90], and reducing bias [156]. As social perspective-taking significantly impacts social interactions, we hypothesize that VSP can improve the behavior of users in virtual and real social interactions. This paper focuses on virtual social interactions, evaluating if VSP improves users' empathic behavior in an interaction with a virtual human.

5.2.2 Avatars Affect Human Behavior

Previous work in using virtual experiences to give the user the perspective of another has focused on placing the user in the avatar of a generic member of a group of which the user is not a member (e.g., embodying a young person with an elderly avatar to improve attitudes towards the elderly [199]). When given an avatar of a person dissimilar to himself, the user takes on the behaviors he expects from a stereotypical member of this dissimilar group (e.g., taller avatars result in more confident behavior; attractive avatars result in more intimate behavior [200]).

While previous work demonstrates that placing a person in the body of a member of a target group causes the person to better identify with the target group, VSP experiences go one step further by placing a person in the specific experience of a member of a target group. In a VSP experience, the user becomes the target, a specific member of a group of people, and relives a specific experience in the target's life. Reliving of a specific

experience has the goal of eliciting reflection on the user's past interactions with both the target and the target's group, in order to improve his future interactions with the class. The VSP experience described in this work seeks to improve medical students' empathic behavior by allowing students to relive their specific interaction with a patient as the patient.

5.3 VSP For Medical Interview Training

To drive our exploration of VSP, we developed a VSP experience based on the interaction between a medical student and Amanda, a mixed reality human (MRH) patient afraid she has breast cancer. The student conducts a medical interview with Amanda using speech and touch (Section 5.3.1). This interaction is recorded from Amanda's perspective (Section 5.3.2) so that the student can relive the interaction as Amanda experienced it (Sections 5.3.3, 5.3.4, 5.3.5).

5.3.1 Simulating a Doctor-Patient Breast Interaction

We chose to use a MRH patient (as opposed to a virtual human patient) because MRH patients incorporate the additional interaction modality of touch and have been shown to elicit more appropriate and empathic behavior than VH patients [99]. The usability of a MRH patient for practicing and evaluating medical students' physical exam skills has been demonstrated [100].

A MRH consists of a life-sized virtual human which is registered to a tangible interface representing the virtual human's body. Users are able to communicate with the MRH through natural speech and touching of the tangible interface; the MRH communicates through pre-recorded speech, gestures (keyframe-based animations), gaze-behavior, and facial expressions.

A MRH patient was created to allow medical students to practice breast history-taking and clinical breast exams. The tangible interface to the MRH patient is a mannequin and physical breast model instrumented with 64 sensors that detect user touch.

The design of the MRH patient is covered in detail in [99]. The user wears a head mounted display (HMD) to view the MRH and a microphone to speak to the MRH. The user's head position is tracked using two Optitrak NaturalPoint cameras and head orientation is tracked by an Intersence InertiaCube2. The user is able to see his touching of the tangible interface in the virtual world. The video stream from a webcam mounted above the mannequin is processed to incorporate the user's hands and the physical breast into the virtual scene. An additional webcam tracks a physical hospital gown attached to the instrumented mannequin, allowing manipulation of the gown as an additional interaction input. A simulation module takes the inputs of user touch, gown manipulation, and user speech and matches these inputs to a database of the MRH's verbal and gestural responses.

5.3.2 Recording an Interaction for VSP

Recording devices log the student's interaction with Amanda (the MRH patient) for later reliving in the VSP experience. Four aspects of the interaction are recorded:

- Amanda's perspective (what she heard and saw)
- Amanda's behavior (what she said and did)
- Clinical breast exam (what she experienced)
- Empathetic challenges (empathy-focused moments)

Amanda's perspective: During the VSP experience, the student must see what Amanda saw and hear what Amanda heard. Hence, Amanda's visual and auditory perspective is recorded by two video cameras (Figure 5-2). Two video cameras are needed because Amanda has two perspectives during the interaction, one for when she is sitting up on the bed and another for when she is lying down. The cameras record the student as he interacts with Amanda, including audio, so that the student can see and hear his conversation with Amanda in the VSP experience.

Amanda's behavior: During the VSP experience, the student must reenact Amanda's speech and movement. Hence, the text of Amanda's speech and movement is logged so that users can be instructed to reenact them during the VSP experience.

Clinical breast exam: During the patient interaction, the student performs a clinical breast exam on Amanda. A clinical breast exam is an uncomfortable, vulnerable moment for patients who fear they may have breast cancer. It requires lying down on a table and exposing one's breast to a doctor for physical examination. During the exam, the doctor searches for potentially cancerous masses by palpating (touching) the patient's breasts. One of the aims of this VSP experience is to help the student understand the feelings of discomfort, vulnerability, and fear that Amanda feels when the student performs the clinical breast exam on her. Hence, the student's exam of Amanda's breast is recorded for later reliving during VSP. As described previously, a webcam mounted above the instrumented mannequin records the user's hands while he touches Amanda's breast. In the VSP experience, the student - reliving Amanda's clinical breast exam - lies down on a medical exam bed and watches as his recorded hands perform a breast exam on his avatar (Figure 5-4).

Empathetic challenges: During the VSP experience, the student relives 3 empathetic challenges, moments where Amanda challenges the student with a statement that requires an empathic response (e.g., Amanda: "I lost my mother to breast cancer. Do you think I could have cancer?"). Hence, the timestamps of empathetic challenges are logged to enable replaying them.

We propose immersing the medical student while replaying the recordings of the interaction to enable sensing what Amanda sensed and reenacting Amanda's behavior. In the VSP, the student sees what she saw, hears what the she heard, touches what she touched, and feels proprioceptively what she felt.

5.3.3 Sensing

5.3.3.1 Seeing and hearing

To see the experience as Amanda saw it, the student wears an Emagin Z800 HMD (800x600, 40-degree fov). The HMD effectively places the student where Amanda was in the virtual world. From this perspective, the student sees the medical exam room where

the original interaction took place, Amanda's body (as his avatar, co-located in the HMD with his real body), and recorded video of himself interacting with Amanda (Figure 5-3).

Head tracking gives the student control over Amanda's viewpoint. The student can look around the room and at his avatar (Amanda), as well as move his viewpoint by moving his head. As Amanda's head position was mostly fixed in the original interaction, the student does not need to move his head much in the VSP experience. Thus, head position is tracked with minimal infrastructure that still provides the visual effects associated with a moving viewpoint (e.g., motion parallax).

Head position is tracked in 2D using a single infrared-viewing webcam and a single infrared-reflective dot on the user's HMD. The webcam is positioned parallel to the student's face so that 2D tracking data can be mapped to vertical and horizontal viewpoint motion. Head orientation is tracked using the Z800's orientation sensor (33hz update rate).

To allow the student to see himself as Amanda did, the recorded video of the student is projected onto a virtual plane. The virtual plane's position was updated to reflect the approximate position of the student in the patient interaction. The video is projected on the plane from approximately the same viewpoint from which it was recorded. As the video was recorded from Amanda's approximate viewpoint, and the student watches the video from Amanda's approximate viewpoint during VSP, the video appears reasonably close to what Amanda saw during the patient interaction. The recorded audio of the interaction is also played back in the VSP experience. This allows the user to evaluate how both his choice of words and nonverbal aspects of speech, like tone of voice and prosody, were perceived by Amanda.

5.3.3.2 Touch and proprioceptive feel

Passive haptics, in the form of a medical exam bed, allow the student to touch what Amanda touched and feel proprioceptively how Amanda felt. During the patient interaction, Amanda sat or lay down on a medical exam bed. Thus, in the VSP

experience, the student sits or lies down on a real exam bed. The real bed is registered to the virtual bed so that Amanda's visual, haptic, and proprioceptive senses are conveyed accurately and in tandem. In terms of haptics, the user can touch the bed and feel its texture and temperature (doctor's offices and exam beds tend to be cold).

The bed also provides subtle proprioceptive cues to match the proprioceptive cues Amanda experienced. While sitting up, the student's legs hang off the bed, and the bed's height provides the feeling of being higher up and looking down on the recording video of the student. While lying down, the student feels the physical sensation of lying down (e.g., gravity acting on the body differently than when sitting, blood flow changes in the body, and arms and legs positioned like Amanda). This enables the student to experience the patient's susceptible position during the physical exam portion of the interaction (Section 5.3.5.3).

5.3.4 Reminders of Virtual Identity

We hypothesize that reminding the user that he is playing the role of the patient during VSP will assist the user in taking the patient's perspective. This is similar to the way that wearing a costume can improve the way an actor plays a part [165], actors need avatars when rehearsing a play in a VE [151], and avatar appearance can affect one's behavior [200]. Thus, throughout the VSP experience, the student is reminded that he is playing Amanda's role.

The student is given Amanda's avatar, which he can see in place of his own body in the HMD. As the HMD had a limited field of view, virtual mirrors are also added to the scene. A mirror on the ceiling allows the user to see his avatar (Amanda) while lying down (Figure 5-4). A green screen behind the video recording of the student is transformed into a mirror, allowing the student to see his avatar while sitting up (Figure 5-5). The student's tracked head movements are linked to his avatar. When the student moves his head, he sees his avatar's head move in the virtual mirrors.

5.3.4.1 Green screen virtual mirror

A green screen is placed behind the student during the patient interaction and replaced with a virtual mirror in the VSP experience. The student sees his avatar's reflection in the mirror, and this reflection changes in real-time with the user's head movement (motion parallax). Virtual mirrors have previously been used to remind users of their avatar [199]. Green screens have been used to create an augmented reality mirror to augment a user's reflection with customized garments [50].

The green screen is transformed into a virtual mirror in two passes, a mirror pass and a green-screen pass. In the mirror pass, the viewpoint of the user is reflected about the mirror plane, the plane onto which the green screen video is projected. The scene is then rendered from the reflected viewpoint into a mirror texture. A clipping plane is placed at the video plane to prevent geometry behind it from being rendered into the mirror texture. In the green-screen pass, green screen texels are replaced by texels from the mirror texture. A fragment shader first segments green texels from the video using a simple color segmentation technique. The technique segments texels which consist of much more green than red or blue ($2G - R - B \geq \alpha$, where α is an experimentally-determined threshold). Segmented green texels are replaced with the mirror texture. While more sophisticated green screen (chroma-key) techniques exist, this simple approach was sufficient for this scenario and lighting conditions.

Subtle cues reinforce the notion that the green screen mirror is a mirror. When the student moves his head, the student sees his avatar's head moving in the mirror, and the mirror image moves to reflect motion parallax. As the green screen is recorded by the cameras during the patient interaction, the parallax cue is strengthened by the student's partial occlusion of the green screen. The parallax cue and head motion also reinforce the notion that the user is in control of Amanda's viewpoint and avatar.

5.3.5 Reenacting Behavior

The student is prompted to reenact Amanda's behavior. Amanda's movement and speech, recorded in the logs, is displayed to the student when it is time to reenact them. For speech, the logged text of Amanda's speech is displayed. For movement, the student's avatar is updated to show the logged movement. Thus, the VSP experience is analogous to acting in a play. The play's script, represented by the logs, provides the student with the direction (speech and motion) needed to play the part of Amanda. The medical student's part in the play was recorded on video during the patient interaction and is played back by the VSP system. By playing back the part of the medical student and prompting the student to play the part of Amanda, VSP enables reliving Amanda's experience of the patient interaction.

5.3.5.1 Speaking

The VSP experience enables reliving the back-and-forth conversation between the student and Amanda. When it is time for the student to speak Amanda's lines, the playback of the medical student's part is paused and a text box appears at the bottom of the student's field of view. The text box displays Amanda's lines for the current point in the patient interaction (Figure 5-6). The student speaks Amanda's lines, and then a human operator resumes playback of the medical student's part. The logs contain the text of Amanda's speech with timestamps, enabling pausing and resuming playback, and displaying Amanda's lines when needed.

5.3.5.2 Moving

To complete playing the part of Amanda, the student reenacts her movements from the patient interaction that are required in a clinical breast exam. These movements are: lying down, sitting up, raising hands in the air, and putting hands on hips.

During the patient interaction, the student instructed Amanda to perform these movements. Thus, during the VSP experience, the playback of these instructions prompts

the student to reenact them. Performing these movements means the student not only reenacts Amanda's behavior but also follows instructions as Amanda did.

Visual cues also remind the student to move as Amanda moved. When one of Amanda's movements is encountered in the logs, the student's avatar is automatically updated to reflect the movement. The student can observe avatar movements in the virtual mirrors to see how to mimic them. For the lying down motion, an additional visual reminder is provided; the student's viewpoint is automatically animated to the lying down position.

5.3.5.3 Reliving a clinical breast exam

While lying down, the student relives the breast exam he performed on Amanda. The student experiences Amanda's physical breast exam through visual pseudo-haptic cues [104, 141]. With visual pseudo-haptic cues, the student's body is not actually touched. Instead, the user sees the visual impact of that touch on his avatar's breast. To enable seeing the impact of his touch, the recorded video of the student's hands touching Amanda's breasts is projected onto the student's avatar. Due to the limited field of view of the HMD, the student could not see the projection while lying down. Instead, the virtual mirror on the ceiling allows the student to see a reflection of his avatar with his recorded hands touching his avatar's breast (Figure 5-4).

5.4 Evaluation of a VSP for Medical Interview Training

A study was conducted to evaluate if VSP affords social perspective taking. We evaluate whether VSP affords social perspective-taking in two ways. First, we examine whether VSP elicits social processes. Social processes must be elicited if the user is to understand how the patient felt as a social actor in this interaction. Second, we examine whether VSP improves use of empathy. If display of empathy improves after VSP, it highlights the ability of VSP to lead to self-directed changes in behavior.

5.4.1 Study Design

Participants conducted breast histories and clinical breast exams of two MRH patients, Amanda and Edna. The first patient interaction serves as a pretest to evaluate a participant's perspective taking skills and empathic behavior. The participant then conducted a VSP experience of the first patient interaction. Finally, the participant interviews a second patient. The second patient interaction is used as a posttest to evaluate how VSP impacts participants' perspective-taking and empathy.

Patient interactions were monitored by one of the investigators, who triggered patient responses if speech recognition failed repeatedly to understand the participants' speech, or to move the conversation forward if the participant asked about a topic (e.g., mental health) for which the patient did not have responses.

5.4.2 Population

Participants were 16 volunteers (10 males, 6 females) from the student and faculty populations at the Medical College of Georgia. One 2nd-year medical student, three 3rd-year medical students, six 4th-year medical students, four residents, and two experienced clinicians participated. All participants had experience conducting breast histories and breast exams of human patients. No participants had experience with life-sized virtual people. Participants' familiarity with video games featuring human or humanoid characters was assessed in a background survey. There was no interaction of video game experience on other measures collected (Section 5.4.4).

5.4.3 Procedure

The study procedure is illustrated in Figure 5-7. Participants completed a patient interaction MRH1, followed by a VSP experience to review the interaction. After the VSP experience, the participant completed a second patient interaction MRH2 to assess if the VSP caused the participant to engage in reflection and self-directed change of his empathic behavior.

In both patient interactions, the participant conversed with the patient for approximately 10 minutes, covering topics of the patient's current complaint and past medical, family, and social histories. This conversation assesses the patient's risk factors for breast cancer. The participant then performed a clinical breast exam of the patient, visually inspecting and touching the patient's breasts. After each stage, a questionnaire was administered.

MRH1: The first patient was Amanda, a 34-year-old Caucasian female with persistent breast pain lasting two months. Amanda was fearful of having cancer, as she had lost her mother to breast cancer two years earlier.

VSP: After the first interaction, the participant reviewed *MRH1* in a VSP experience as described in Section 5.3. To encourage participants to reenact Amanda's actions, participants were instructed to mimic Amanda physically (e.g., "lie down when the patient does") and verbally (e.g., "read the patient's lines when they appear on screen").

MRH2: The second patient was Edna, a 53-year-old Caucasian female who had recently found an area of thicker breast tissue. Edna was also fearful that she may have breast cancer, as she had experienced her mother and a friend's bouts with breast cancer, both ending in mastectomies and depression.

5.4.3.1 Patient interactions

Both patient interactions were deemed equivalent in both medical content and opportunities to express empathy. Each provided three main opportunities for expressing empathy.

After the participant conversed with the patient for 10 minutes, the participant began a clinical breast exam of the patient. The breast exam began with the participant asking the patient to remove her gown (bare her breasts) in order to visually inspect her breasts for abnormalities. When the patient was asked to remove her gown, she expressed that she was nervous and shy about being naked in front of the doctor.

- **Amanda:** "I guess that's ok. I'm kind of shy about taking off my clothes."
- **Edna [anxiously]:** "Oh ok."

The second empathetic moment occurred as the participant removed the patient's gown in order to begin palpating (touching) the patient's breasts. As the participant opened the patient's gown, the patient verbally expressed that she wanted to stop the exam because she was fearful that the exam might find cancer:

- **Amanda:** "Wait! I'm scared you might find something bad. What if you find cancer?"
- **Edna:** "Wait! Before you do the exam, tell me what is the chance this is something other than a normal breast change?"

The third empathetic moment occurred as the participants explained his findings and next steps in the patient's treatment. Both patients had a mass, so the findings of the participant should be that a mass had been found in the patient's breast. The next step in treatment was for Ms. Jones to receive a mammogram (because she had never received one) and for Ms. Jacobs to have a biopsy of the mass (because she recently had a mammogram which failed to highlight the mass). Each patient expressed a fear of the treatment:

- **Amanda:** "Do I really have to get a mammogram? My mom was fine, then she had a mammogram and they said she had cancer, and all of a sudden she was really sick!"
- **Edna:** "A friend of mine had a biopsy and by the time they went back to do the surgery, the cancer had spread. I had to watch her and my mom go through radiation and mastectomies. It was really hard for them to deal with. Could this happen to me if I have a biopsy?"

Both of the patients' stories of fear demonstrate an intertwining of incorrect medical knowledge with complex psychological issues of fear and loss. Responding to these fears appropriately requires understanding the patient's perspective. The participant should handle this situation by both empathizing and educating, without demeaning the patient or her concerns. Any improvement in empathy from the first to second patient interaction is seen as being due to the social perspective-taking afforded by the VSP experience administered between the two patient interactions.

5.4.3.2 VSP experience

Before beginning the VSP experience, an investigator explained to the participant that he was going to switch places with Amanda, the patient he just interviewed, and experience the interview as Amanda. As Amanda began the patient interaction by sitting on the corner of a medical exam bed, the participant began the VSP experience by also sitting on the same corner of a real medical examination bed. The participant then donned an HMD to view the virtual world through Amanda's perspective. The participant was instructed to look down to see Amanda's body in place of his own. The investigator also asked the participant to mimic Amanda's speech and movements as described in Section 5.3.5.

After these instructions, the participant spent 10 minutes reliving the patient interaction as Amanda. To accustom the participant to the VSP experience, the student relived the first minute (greeting, patient's complaint) of *MRH1*. Then the participant relived the empathetic moments of *MRH1*. Rather than relive each empathetic moment in isolation, the participant relived a minute long period starting 30 seconds before the empathetic moment. Reliving this full minute provided conversational context for the empathetic moment.

5.4.4 Measures

Measures assessed participants' perspective-taking, use of empathy, and copresence (an indicator of social processes with VHs). The same questionnaire was given after each patient interaction and the VSP experience. The questionnaire given after the first patient interaction is denoted Post-MRH1; after the VSP is denoted as Post-VSP; and the questionnaire given after the second patient interaction is denoted as Post-MRH2. One participant was called to attend to human patients before completing the Post-MRH2 survey, and another participant left items blank, so the population used for comparisons of Post-MRH1 and Post-VSP is $n = 15$; for comparisons of Post-VSP and Post-MRH2 the population is $n = 14$. Another questionnaire given after VSP assessed participants'

perceived benefit of VSP. In this section, means and standard deviations are reported in the form of M $\pm$SD. Statistical significance was set at $\alpha < 0.05$.

5.4.4.1 Empathic behavior and perspective-taking

Empathic behavior was measured with a questionnaire (Table 5-1) consisting of ten 5-point Likert items adapted from empathy subscales of validated instruments used in medical education: the Jefferson Scale of Physician Empathy [81] and the 4-Habits Coding Scheme [101]. Reliability of this questionnaire was high (Cronbach's $\alpha = 0.73$).

Perspective-taking was measured with a questionnaire (Table 5-1) of four 5-point Likert items also adapted from [81, 101]. The perspective-taking questionnaire has high internal consistency (Cronbach's $\alpha = 0.91$).

5.4.4.2 Copresence

The copresence instrument used after the patient interactions is that of Bailenson et al [16]. This instrument was also adapted to measure copresence in the VSP experience. Specifically, three 5-point Likert items were adapted: (1) During my time in the role of the patient, I made eye contact with the doctor. (2) During my time in the role of the patient I felt like I was talking to another person. (3) I felt that the doctor was aware of my presence.

Reliability of the VSP copresence instrument was high (Cronbach's $\alpha = 0.83$). Enforcing that the construct this instrument measures is indeed copresence, a large correlation was found between this instrument and the copresence instrument used in the post patient-interaction questionnaire. Post-VSP copresence and Post-MRH1 copresence were significantly correlated (Pearson's $r(13) = 0.65, p < 0.05$). Post-VSP copresence and Post-MRH2 copresence were also significantly correlated (Pearson's $r(13) = 0.81, p < 0.0005$).

5.4.5 Results

5.4.5.1 VSPs elicit reflection and self-directed change

During the VSP, participants reported reflecting on their use of empathy and perspective-taking in the first patient interaction. Participants' self-ratings of their use of perspective-taking significantly decreased from Post-MRH1 to Post-VSP (Post-MRH1: 3.98 ±0.55; Post-VSP: 3.67 ±0.61; F(1,15) = 5.6, p ¡ 0.05). Additionally, most participants' ratings of their use of empathy decreased from Post-MRH1 to Post-VSP (9 decreased, 3 increased, and 2 did not change), a close-to-significant difference (Post-MRH1: 3.71 ±0.72; Post-VSP: 3.49 ±0.89; F(1,15) = 4.53, p = 0.055). These results demonstrate that the social perspective taking afforded by the *VSP* experience elicited self-reflection by participants. Reflecting on their use of empathy and perspective taking in the patient interview, participants realized that they had not adequately taken the patient's perspective or empathized with the patient during the *MRH1* patient interaction.

The reflection afforded by VSP allowed participants to engage in self-directed change of their behavior in the second patient interaction. This is evidenced by a self-reported improvement in participants' use of empathy and perspective taking in the second patient interaction. Participants rated themselves as improving significantly in their use of empathy (Post-VSP: 3.49 ±0.89; Post-MRH2: 3.88 ±0.44; F(1,14) = 7.0, p ¡ 0.05) and trended towards an improvement in perspective taking (Post-VSP: 3.67 ±0.61; Post-MRH2: 3.95 ±0.45; F(1,14) = 3.7, p = 0.076). The self-rated improvement in empathy and perspective-taking indicates that participants tried to improve their use of empathy and perspective-taking in the second patient interaction. This self-directed change is brought about by the reflection afforded by taking the patient's perspective during the VSP experience.

5.4.5.2 VSP affords social processes and interaction

While playing the role of the patient in the VSP experience, participants reported a high degree of copresence with the doctor (the recording of themselves from the patient

interaction). Eliciting a high copresence indicates that VSP is an inherently social experience that affords social processes and interaction.

Eleven of 14 participants reported making eye contact with the doctor. Eye contact is a social process and important component of nonverbal communication. Although the participant could not actually make eye contact with the doctor (the doctor's face was covered by an HMD), the responses indicate participants were socially present with the doctor and attempted eye contact to communicate nonverbally and act out the social convention of making eye contact when speaking and listening as they would in a conversation with patients.

A significant majority of participants felt they were talking to another person during the VSP experience, indicating that VSP supports social interaction through verbal communication as well. 8 of 14 participants felt the doctor was aware of their presence.

5.4.5.3 Observations

During the VSP experience, the majority of participants mimicked the patient's behavior as instructed. Just as the patient did, participants lay back on the exam table (15 of 15), raised their hands over their head (9 of 12), put their hands on their hips (8 of 10), and placed their left arm under their head when they lay down (7 of 14). We propose that mimicking patient behavior is an indicator of *role presence*, or the extent to which the user transports himself into the role of the patient.

Participants commented on what they learned from VSP and what aspect of their patient interaction VSP motivated them to improve. The areas most commonly identified as needing improvement were a need to better express empathy (3 participants), a need to display more confidence (3 participants), and a need to better address patient's concerns, including the patient's fears (4 participants). A common realization of participants was that they look different in a patient interview than they thought, e.g., "What I thought I looked like in interviewing patients is much different than I actually do." Participants also commented how the VSP experience would change their future interactions with

human patients. Participants agreed that "Reviewing the interview and exam changed how I will address patient fears" (10 agree / 1 disagree), that "The review session increased my understanding of patient's emotions in sensitive situations" (12 agree / 0 disagree), and "After reviewing my exam of the patient, I expect to be more empathetic in future intimate exam scenarios." (10 agree / 2 disagree). Participants perceived VSP as improving their understanding of the patient's perspective and reported that VSP would improve their behavior in future interactions with human patients.

5.4.6 Discussion

5.4.6.1 VSP elicits social processes

VSP builds upon previous uses of virtual reality for facilitating and training social interaction. In our VSP experience, users had a high sense of copresence with their recorded self. Additionally, the VSP experience had the appearance of a conversation, with the participant speaking the patient's words and the recording of the participant speaking back. Participant responses indicate they felt like they were having a conversation with another person.

VSP is successful at facilitating social perspective-taking and self-directed change precisely because it elicits social processes. Without social processes, such as copresence and conversational interaction, participants would have difficulty understanding what it was like for the patient to interact socially with the participant.

The social processes elicited by VSP also show that participants felt immersed in the experience. Despite the artificial nature of the patient and VSP interactions, participants were willing to immerse themselves socially in the VSP interaction.

5.4.6.2 VSP motivates self-directed change

In improving empathy, VSP motivated self-directed change in empathy skills. VSP helped participants identify poor empathy skills in the first patient interaction and motivated them to improve those skills in the second patient interaction. This also extends

prior work in after-action reviews (AARs) of human-virtual human experiences [147] by demonstrating that change in behavior with VHs can be elicited through AARs.

5.4.6.3 Seeing the world through unfamiliar eyes

This paper's broader impact on VR is in enabling seeing the world through unfamiliar eyes. In our VSP experience, several male doctors were able to experience a female patient's clinical breast interview and exam. The likelihood of a male doctor receiving a breast exam, and understanding the patient's physical and emotional perspective of the experience, is low. Thus, VSP provides an opportunity for users to see the world in a way they might never see it and learn from that experience.

5.5 Conclusions and Future Work

This paper proposes virtual social perspective-taking (VSP). VSP leverages virtual environments to help users understand and learn from the experiences of others. Three guidelines for VSP were proposed: (1) replace the user's senses with the recorded senses of another person, (2) encourage the user to reenact the other person's behavior, and (3) remind the user that they are another person. These guidelines were applied to an interaction between a medical student and a virtual human (VH) patient, enabling the student to relive his interaction with a VH - *as the VH*.

Reliving the interaction as the VH effectively allowed seeing what it is like to talk to oneself. This provided healthcare students with an important educational experience of reflecting on how they are perceived by patients, and on what it was like to be the patient. This led students to self-identify skills in need of improvement and change behavior. These responses are difficult to realize with current approaches, such as group and instructor video review. Thus this work highlights the expansion of VR to support the education of social perspective-taking.

While our results indicate VSP elicited perspective-taking, we hypothesize more impactful perspective-taking could be elicited by allowing the user to *think and decide* how to reenact the VH's behavior. Currently, users are told what to say and shown how to

move, so they have little need to consider the VH's perspective to play her role. In future work, we will ask the user to speak what they *think* the VH would say. We expect this will encourage the student to think about the VH's perspective before speaking and result in greater use of empathy in future interactions.

We also propose using VSP to allow the student to relive the VH's interaction with an empathy expert (e.g., an oncologist with identified expert experience in interacting with cancer patients). We expect this to improve social perspective-taking by allowing the student to contrast a patient's perspective of the expert with the patient's perspective of himself.

Lastly, we believe the next step for VSP is enabling VSP of human-human (H-H) social interactions. Implementing the three VSP guidelines for H-H interactions would be challenging. If successful, users would be able to relive and learn from the experiences of the people they interact with everyday.

When the patient expressed fear, I felt sorry for the patient
I did not pay attention to the patient's emotions when interviewing and examining her
I did not feel that it was important for me to gain an understanding of how the patient felt about her mother's experience with cancer
I felt that I made an emotional connection with the patient
I encouraged the patient to express her emotions
I accepted and/or validated the patient's feelings
I displayed little interest or concern to the patient
I made little or no attempt to explore the patient's feelings
I legitimized the patient's ideas and feelings
I demonstrated appropriate non-verbal behavior
I tried to put myself in the patient's position to help understand her fear
I could not understand why the patient was fearful
It was difficult for me to view things from the patient's perspective
I understood the patient's perspective

Table 5-1. Empathy (top) and perspective-taking (bottom) questionnaires.

Figure 5-1. Medical student converses with virtual patient (Top), then relives conversation *as the patient* (Bottom).

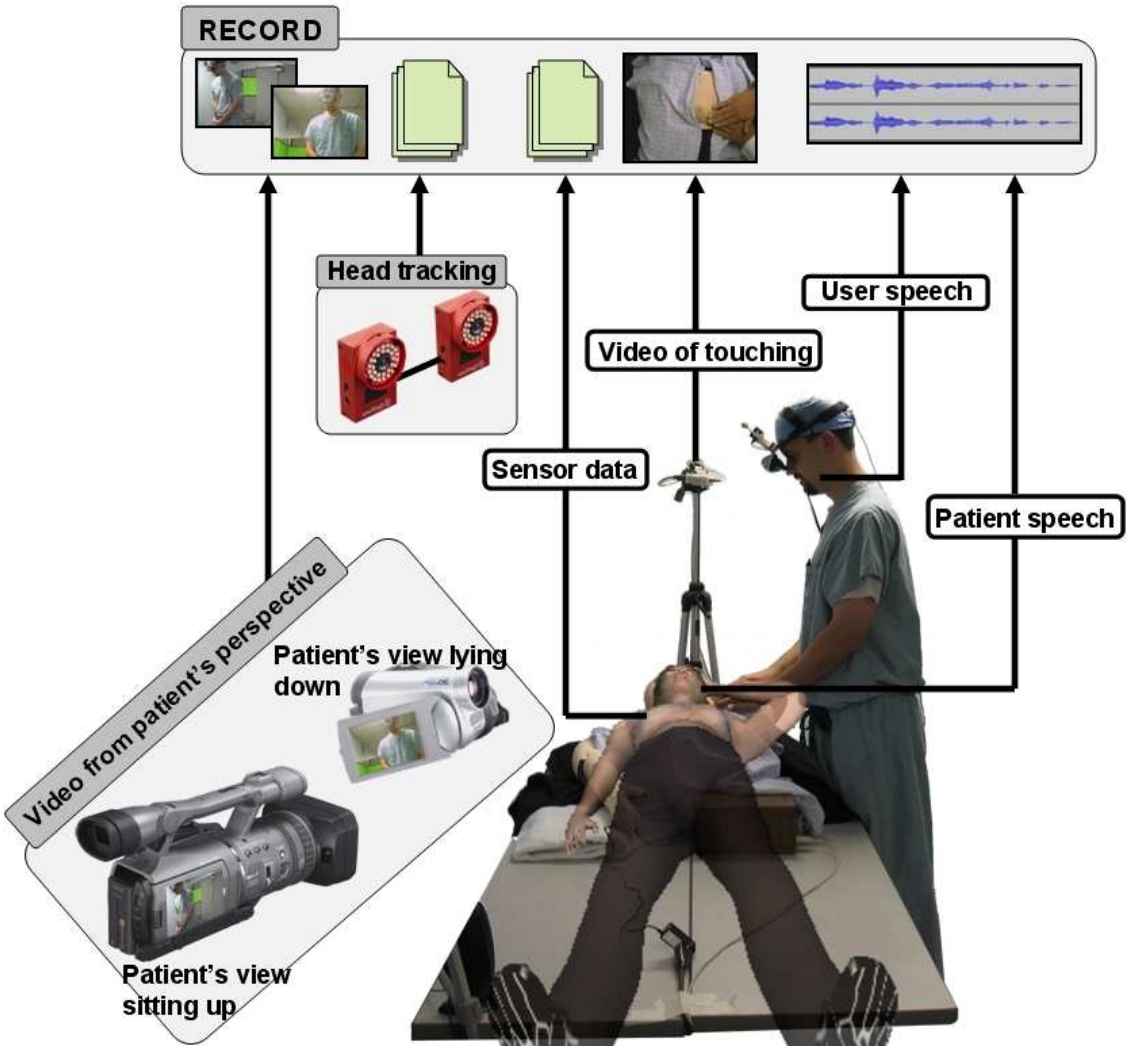


Figure 5-2. Logging of patient interaction for VSP

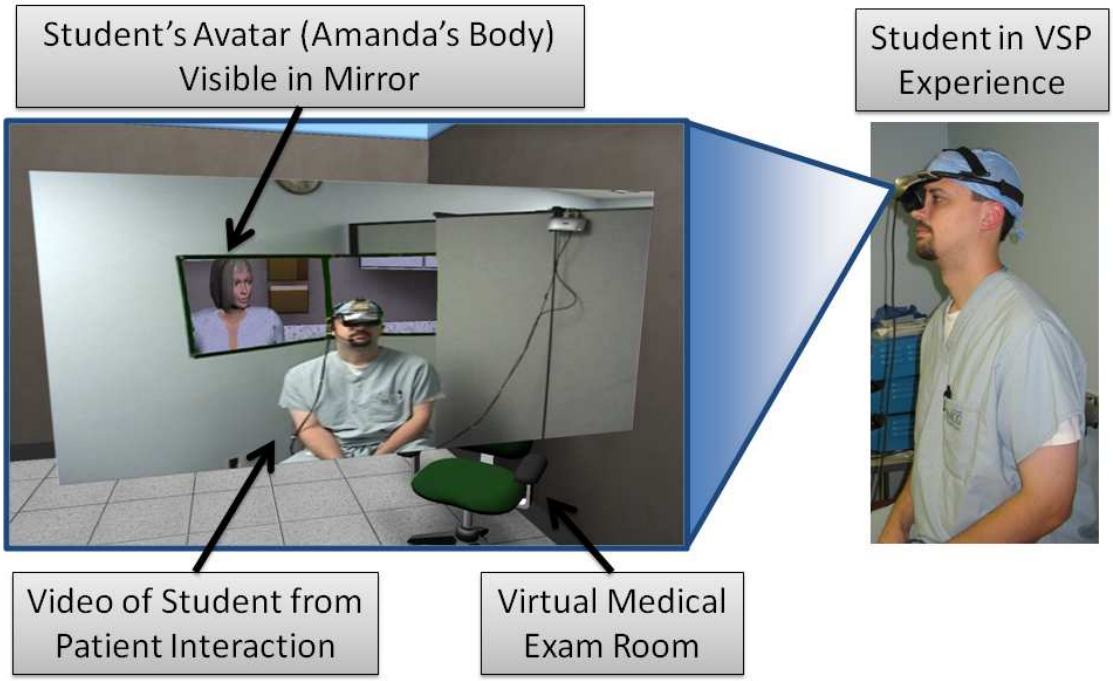


Figure 5-3. During VSP, the student sees the exam room, his avatar (Amanda's body), and video of himself talking to Amanda from Amanda's perspective.

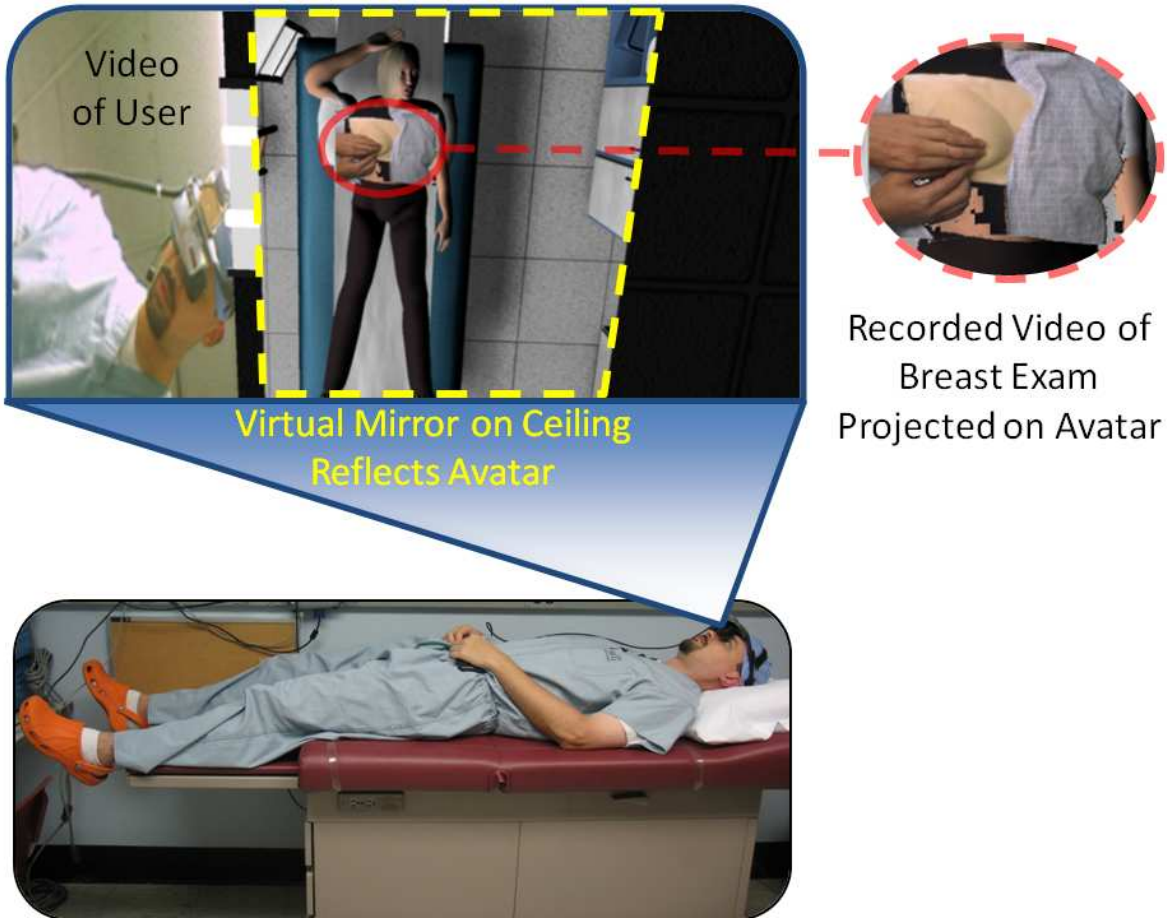


Figure 5-4. In the VSP experience, the student, playing the role of Amanda, relives the breast exam he performed on Amanda.

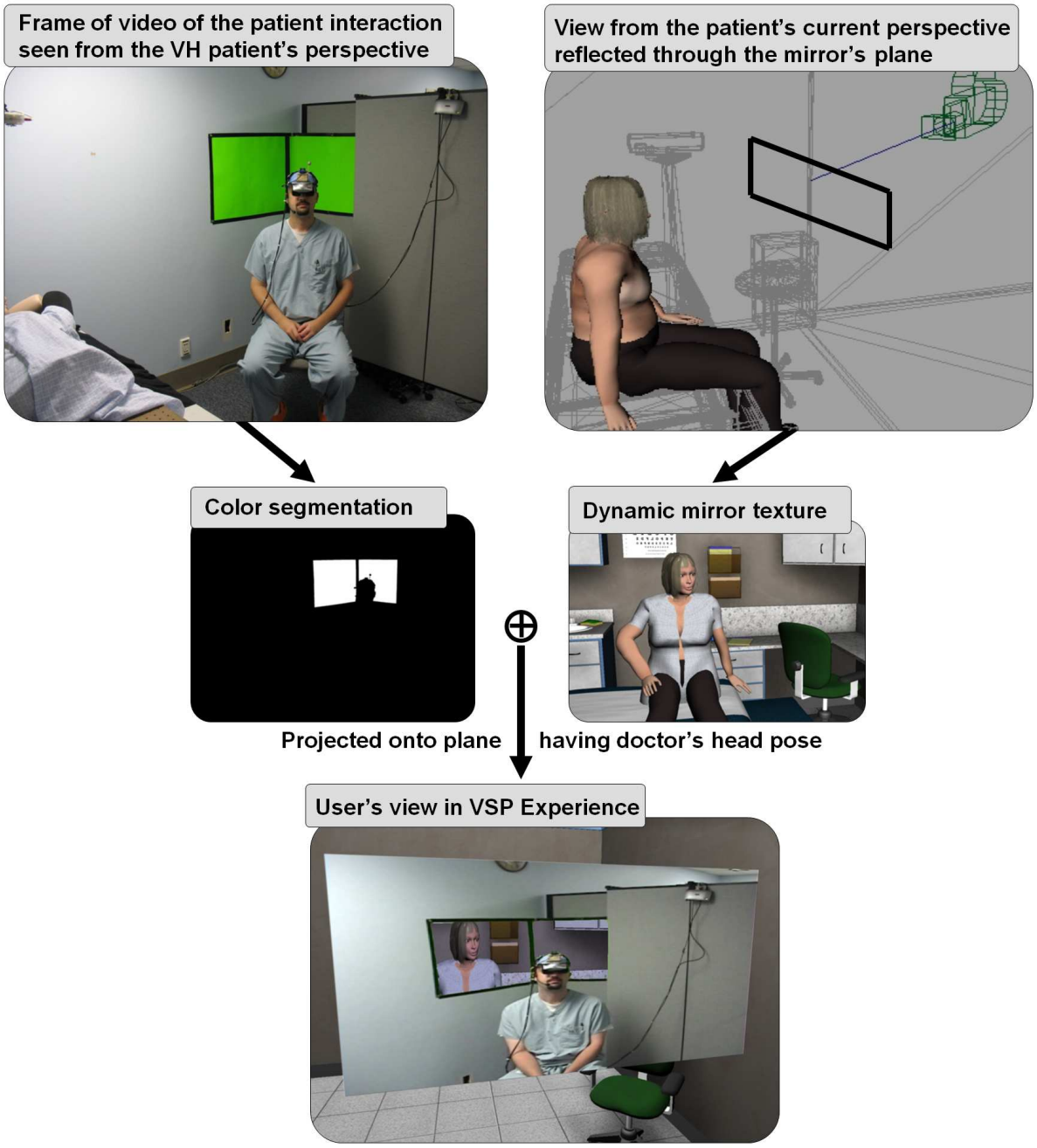


Figure 5-5. Transforming the green screen into a virtual mirror.

Timestamp	Virtual Human Speech
0:16	I'm ok I guess, I'm kind of nervous though
9:33	I've never had cancer
9:47	yes
10:06	I don't know, I mean don't mammograms hurt?
10:33	do I really have to get a mammogram? I mean my mom was fine then she had a mammogram and they said she had cancer and then all of the sudden she was really sick
10:45	yeah
11:49	thank you doctor
11:54	ok I'll see you then
12:01	--INTERACTION COMPLETE|-

Figure 5-6. VH speech logged for later display in the VSP experience.

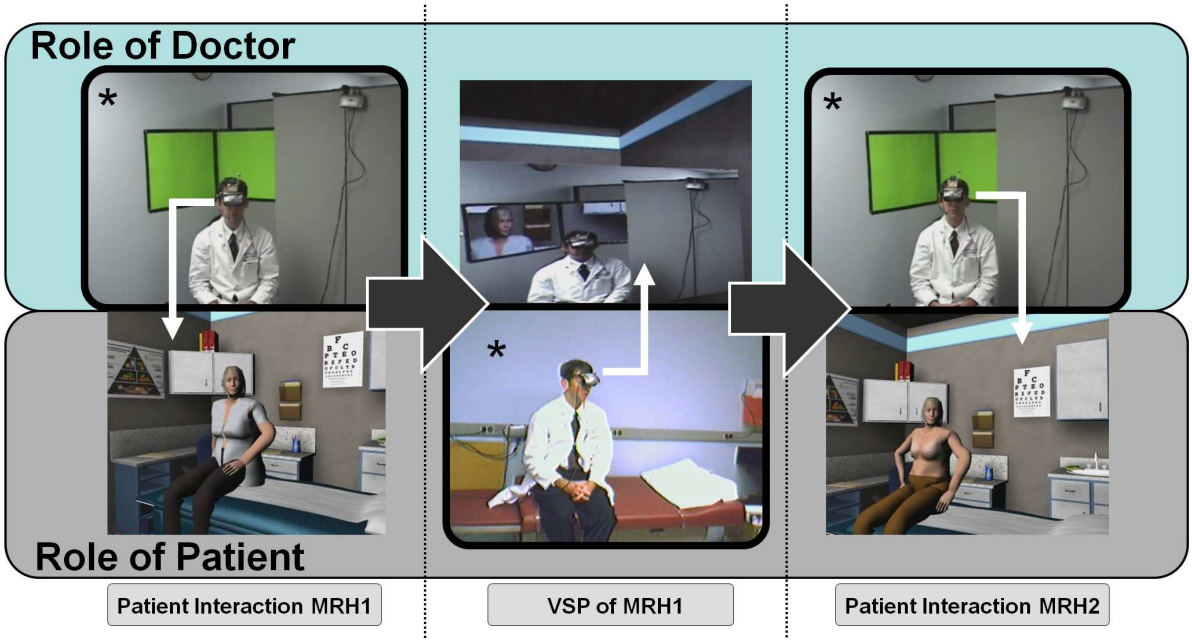


Figure 5-7. Study procedure: bordered images indicate role (patient or doctor) participant played at each stage.

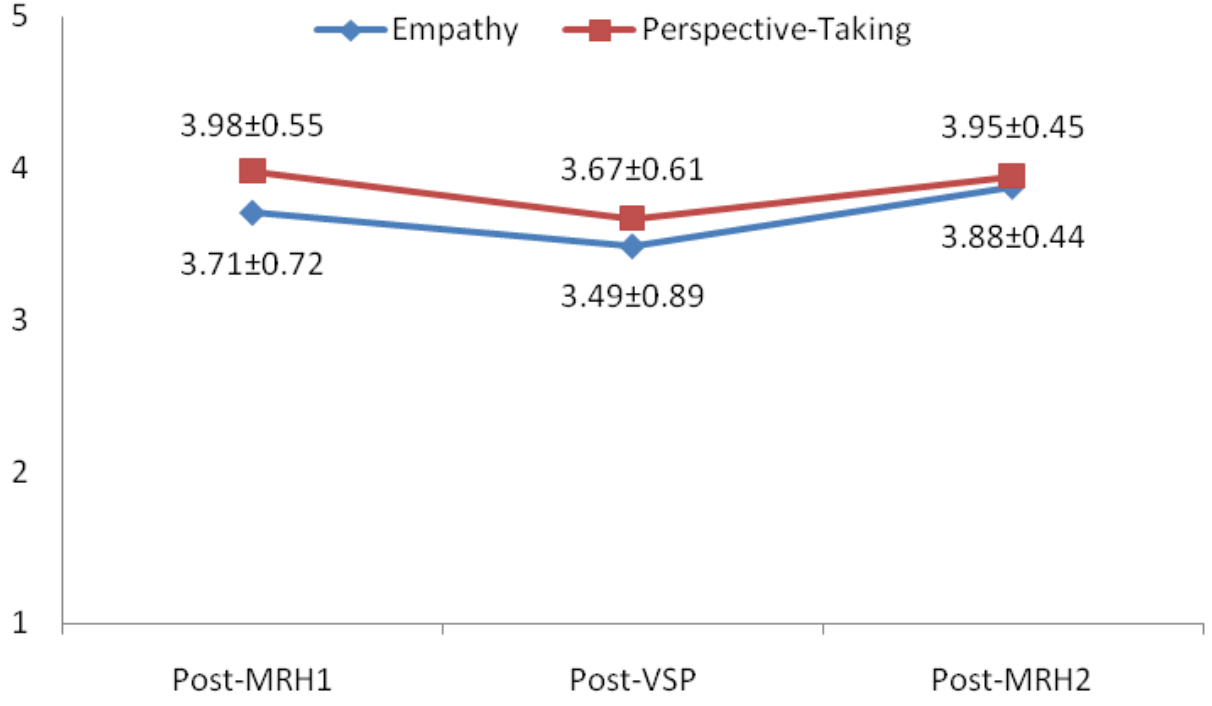


Figure 5-8. Empathy and perspective-taking improved after VSP, indicating participants tried to improve empathy and perspective-taking in the 2nd patient interaction.

CHAPTER 6
IPSVIZ^N: AGGREGATE AFTER-ACTION REVIEW FOR VIRTUAL HUMAN EXPERIENCES

This chapter describes IPSViz^N, a desktop-based, after-action review (AAR) system for groups of interactions with virtual humans (VHs). IPSViz^N leverages visualization to present overviews and details of aggregate H-VH experience data. The visualizations allow users to uncover trends and outliers in interpersonal communication between humans and virtual humans. A preliminary evaluation explores the benefits of conducting aggregate AARs of H-VH experiences.

Personal contributions: I developed IPSViz^N in its entirety and solicited feedback from domain experts using IPSViz^N.

Collaborators: D. Scott Lind, Juan Cendan, Diane Beck, Carole Kimberlin, Adeline Deladisma and Toufic Imam contributed medical and pharmaceutical education expertise that influenced the design of IPSViz^N.

Relevance to thesis: IPSViz^N is the last of three AAR systems for H-VH experiences developed as part of this dissertation. IPSViz^N demonstrates the analysis of groups of human-virtual human interactions using visualization techniques, and shows how this analysis leads to reflection on, new perceptions of, and changes in domain-specific interpersonal skills.

6.1 Introduction

This chapter describes IPSViz^N, a tool that addresses the challenge of rapidly analyzing and uncovering trends and outliers in large numbers of training interactions with virtual humans. IPSViz^N generates visualizations that summarize and highlight outliers in the aggregate behavior (verbal and nonverbal) of users who interacted with a virtual human. Details about individual users and groups of users is facilitated through filtering, querying, data rescaling, and semantic zooming [138]. The utility of IPSViz^N is demonstrated by using it to analyze interpersonal simulations - interactions between humans and virtual humans for interpersonal skills training.

6.1.1 Motivation: Feedback and Evaluation of Interpersonal Skills

The traditional approach to evaluate interpersonal skills is to observe trainees interacting with a partner. Evaluations are conducted by the trainees, the partner, or by an external observer who does not participate directly in the interaction (e.g., an instructor or peers). Evaluations can take place during the training session or afterwards on video. Often, the trainee receives feedback based on the evaluation and incorporates lessons learned into future interpersonal interactions.

This traditional approach has several limitations for evaluating interpersonal skills, including 1) the time investment required for evaluation, 2) the lack of availability of interaction participants, and 3) the "hidden nature" of important interpersonal skills.

- **Time investment:** Reviewing and evaluating interpersonal interactions takes time. Evaluators must commit time to watching most (if not all) of an interaction to grade it. This time commitment scales linearly with the number of interactions; the more interactions to evaluate, the more time needed for evaluation. For example, an interpersonal skills instructor with 180 students would find it difficult to thoroughly review and evaluate all student interpersonal interactions, especially since each student may conduct multiple interactions over the course of the semester

- **Availability:** Interaction partners are often a limited resource. In many cases, they must be trained to present a specific scenario, and must be paid for their duties. Thus, they are usually available in small numbers (only a handful of well-trained individuals) for limited times (once a week). The lack of availability means trainees will experience infrequent interpersonal interaction and thus receive infrequent, insufficient feedback. Furthermore, the significant time investment the evaluator must invest to evaluate the interactions further limits feedback to trainees.

- **"Hidden" interpersonal skills:** Many interpersonal skills, such as nonverbal behavior, are not explicitly visible or understandable to evaluators. Evaluators must be trained to see, evaluate, and provide feedback on them. Without training, an evaluator may be able to observe an interaction and provide a high-level evaluation ("the trainee was aloof"), but it would be more difficult for the evaluator to highlight low-level details (when aloofness was displayed and what aspects of verbal and nonverbal behavior were responsible for portraying aloofness). Note that the issue of hidden interpersonal skills is related to the time investment problem, as most if not all of an interaction must be reviewed to get a holistic sense of both the high-level ("aloofness") and low-level ("poor posture coupled with a low tone of voice at the three minute mark") feedback.

6.1.2 Motivation: Virtual Humans for Interpersonal Skill Training

Recently, interactions with virtual humans (VHs) have emerged as a viable tool for interpersonal skills training and evaluation. Virtual humans are good for interpersonal skills training because the interpersonal skills a person demonstrates in an interaction with a virtual human are similar to the interpersonal skills that same person demonstrates in an interaction with a real human [89, 145]. By extension, evaluating interactions with a virtual human should provide insight into interactions with real humans.

In addition, virtual humans can address the multiple challenges posed by evaluating interactions with real humans:

- **Time investment:** Interactions with virtual humans can be logged and evaluated computationally to provide rapid insight into interpersonal skills. Lengthy review of interactions can be replaced with focused reviews of specific skills and moments highlighted by the computational evaluation.
- **Availability:** Virtual humans can be made available at all times for interpersonal interactions. Thus, users could practice an interaction whenever they have time, and then receive a computational evaluation of that interaction afterwards. The increased availability of virtual humans also tightens the feedback loop. Users can interact with a virtual human, get a quick evaluation of that interaction, conduct another interaction, be reevaluated, and so forth, all within a short time window (e.g., a few hours). The extra practice and feedback means users 1) improve skills more rapidly, and 2) can track progress more readily by comparing how computational evaluations evolve over time.
- **"Hidden" interpersonal skills:** By leveraging interaction logs, hidden interpersonal skills can be made explicitly visible to an evaluator. The log can include VH body motions and 3d tracking data of the user (nonverbal behavior), as well as recordings of user and VH speech (verbal behavior). This data can be processed and summarized computationally to provide rapid insight into hidden behaviors without requiring extensive training of evaluators. This also saves significant time for evaluators, since the processing and summaries can quickly highlight important aspects of the interaction without the need to review it in its entirety.

6.1.3 The Interpersonal Scenario Visualizer

In previous work, the Interpersonal Scenario Visualizer (IPSViz) [147] was proposed to support evaluation of an interaction with a virtual human. IPSViz transforms the log of an interaction between a human and virtual human into a set of interactive visualizations.

The visualizations enable evaluation of the interaction in ways that would be difficult or impossible to do with interactions with real humans. These include analysis of user and VH verbal and nonverbal communication at multiple levels of detail, rapid review of important moments in the interaction, and reviewing the interaction from multiple perspectives including through the eyes of the virtual human and an external observer.

In a user study, IPSViz users reviewed their interaction with a virtual human to learn about their real-world interpersonal skills. After using IPSViz to review their own interaction with a virtual human, participants reflected on and changed self-evaluations of their interpersonal skills. Several participants highlighted ways they would change behavior in future interactions with real humans as a result of the review with IPSViz. The skills highlighted by students were skills that they had not considered previously in their interactions with real humans. Thus, the study confirms the benefits gained from using an interaction with a virtual human as the basis for interpersonal skills feedback: rapid, frequent review of "hidden" interpersonal skills (verbal and nonverbal communication) in ways that would be difficult or impossible to do through interactions with real humans.

6.1.4 IPSViz^N

This chapter builds on IPSViz with IPSViz^N, a visualization tool for evaluating *many (or N)* interactions with virtual humans. Like IPSViz, IPSViz^N enables rapid, frequent review of domain-specific interpersonal skills (verbal and nonverbal communication) in ways that would be difficult or impossible to do through interactions with real humans. However, IPSViz^N differs from IPSViz in that it leverages information visualization techniques to enable evaluation of the aggregate interpersonal skills of *groups*.

6.2 Analysis Tasks

Two important analysis tasks for groups of interactions with virtual humans are:

1. Discovering trends and outliers (deviations from trends) in groups

2. Comparing an interaction or group of interactions to another interaction or group of interactions

These analysis tasks are difficult or time-consuming to perform using existing methods for the following reasons:

- **Linear review:** With existing analysis methods such as video review, trends and outliers can only be discovered through linear review of many interactions. Linear review means reviewing interaction data in its entirety, in the order in which it was collected. Linear review is necessary because the reviewer often does not know what to look for in the interactions. Entire interactions must be reviewed to identify aspects of interest and insure that nothing important is missed. Unfortunately, linear review can be tedious and require a significant time investment. Furthermore, this time investment grows linearly with the number of interactions to be analyzed.

- **Tracking trends and outliers:** Linear review requires tracking and updating trends and outliers as the interactions are reviewed. Keeping track of trends and outliers across many interactions can be a difficult cognitive task. Outliers, by definition, are rare and might be difficult to identify and keep track of when linearly reviewing a large set of interaction data.

- **Interpersonal behavior:** As users communicate with VHs through verbal and nonverbal behavior, verbal and nonverbal behavior should be analyzed in H-VH interactions. Unfortunately, traditional analysis methods, such as video review, do not provide the level-of-detail needed to analyze the subtleties and relationships in verbal and nonverbal behavior.

For example, in a typical, 10-week, medical communication skills course, 180 students would conduct eight 10-minute interviews with virtual human patients. That amounts to 240 hours of linear video review total, or 24 hours of video review a week, that the instructor would have to conduct to gain a holistic sense of whom the good and bad students are (outliers) and overall trends in student performance. Performance in a medical interview is measured in terms of the following technical (e.g. diagnosis, treatment, and physical examination) and interpersonal medical skills [35, 167]:

- **Information gathering:** Gathering complete and accurate information from the patient for diagnosis

- **Rapport:** Building rapport to encourage the patient to speak freely and honestly

- **Empathy:** Expressing empathy appropriately when the patient is afraid or when delivering bad news
- **Diagnosis and treatment:** Explaining diagnoses and treatment steps clearly and honestly

To properly evaluate the class of 180 students on these interpersonal skills, the instructor needs to track, summarize, and compare the students' and virtual humans' behaviors across 240 hours of video. These behaviors include:

- **Information gathering:** What questions the students asked the virtual human and how the virtual human responded.
- **Rapport:** What verbal and nonverbal behavior (e.g., tone of voice, posture, and gaze behavior) participants used to create a sense of openness and comfort throughout the interview.
- **Empathy:** If the students expressed empathy when necessary (when the virtual human needed it) using appropriate verbal and nonverbal behavior.
- **Diagnosis and treatment:** If the students explained diagnoses and treatment steps and what words the students used to do so.

IPSViz^N solves the challenges inherent in tracking and identifying trends and outliers in interpersonal skills across large numbers of H-VH interactions. The challenges are addressed by capturing the verbal and nonverbal behavior used in H-VH interactions and using information visualization techniques to provide users with rapid insight into them.

1. **Capture verbal and nonverbal behavior:** Sensors, such as microphones and 3D trackers, capture a user's verbal and nonverbal behavior, and logging facilities capture the virtual human's behavior in detail. This provides a computational description of domain-specific interpersonal skills for further processing and analysis.
2. **Generate overview visualizations:** The captured interactions are computationally summarized and overview visualizations are generated from the summaries. Overview visualizations allow users to see what happened in the interaction at a glance without the need for linear review of large amounts of data. This enables rapid identification of group trends in the interactions. Furthermore, outliers in H-VH interactions become easier to discover because they stand out from overview visualizations (e.g., an unusual spike in blood pressure in an otherwise smooth curve).

3. **Provide interactive analysis tools:** Interactive tools allow users to filter, rescale, query, and zoom into data presented by the overview visualizations. Filtering allows users to rapidly compare subgroup trends on different interpersonal skills by selectively displaying each subgroup and skill. Rescaling allows users to compare interactions on different time-scales. Querying allows users to map unusual data points (outliers) back to individual interactions or users.

To provide an initial evaluation of this approach, IPSViz^N was applied to a simplified version of the medical education scenario described previously (evaluating a class of students). Interactions between 29 medical professionals and a virtual human complaining of breast pain were captured. Behaviors captured included the words spoken by the user and virtual human, meta-topics which summarize these words, and nonverbal behavior (tone of voice, posture, and facial expressions) from video of the interactions. These behaviors had specific relevance to this medical scenario, and allowed for analysis of trends and outliers in information gathering, rapport, empathy, and diagnosis and treatment. Using IPSViz^N, visualizations of the captured behaviors were then provided to medical educators. By reviewing and interacting with the visualizations, the educators identified trends and outliers in user behavior that would be difficult or impossible to identify otherwise.

6.3 Virtual Human Patients

6.3.1 Motivation

Interactions with virtual human patients provide a real-world, validated platform [89] for exploring the underlying questions posed by this chapter:

- **Can one evaluate a group's real-world interpersonal skills by evaluating their interactions with virtual humans?** Previous research shows that VH patient interactions are validated for demonstrating real-world interpersonal skills [89]. Furthermore, interactions with VH patients are similar to interactions with simulated patients (human actors). Lastly, work applying after-action reviews to H-VH experiences [146, 147] has found that user can self-evaluate their interpersonal skills by reviewing their interaction with a virtual human patient. These results provide evidence and motivation for evaluating *group* interpersonal skills through interactions with virtual humans.

- **Can we generate visualizations that 1) aggregate logs of H-VH interactions to 2) provide insight (trends and outliers) into the interpersonal communication of groups that 3) would be difficult or impossible to provide otherwise?** Fundamental information visualizations principles, such as Shneiderman's mantra [161], provide guidelines for gaining insight into large datasets through information visualization. In addition, several compelling visualizations of interpersonal communication have been developed [5, 20, 22, 46, 93, 160, 186, 187, 191]. Lastly, previous work has shown visualizations of individual interactions with VH patients can be visualized to provide insight into interpersonal skills [147]. These visualizations provided insight that were difficult or impossible to receive through practice interactions with real humans.

- **What kinds of insight do trainees and trainers need to learn from H-VH experiences that they are not learning otherwise?** Those who interact with the virtual human patient (students) want to know how they fared on medical communication skills both in comparison to their previous interactions and in comparison to their peers. Those who administer or design VH patient interactions (medical educators) want to know if the interactions meet pedagogical goals across groups of users, whether students are learning from the interactions, whether there are trends among subgroups of the student population (e.g., do men express empathy as much women?), and which students are doing poorly and need remediation. These use cases all fall under the general case of identifying and comparing trends and outliers in individual and group behavior.

6.3.2 Interaction

In this study, the interactions between a group of medical professionals (n=29) and a virtual human patient named Amanda were captured for later review in IPSViz^N. Amanda had been experiencing breast pain for about a month. She has a family history of breast cancer and was concerned her breast pain meant she had cancer. Much like a medical interview with a real patient, each medical professional (MP) asked Amanda a series of questions to learn about the history of her pain, her family history, and other relevant medical information. After collecting this information, the MP conducted a physical breast exam on Amanda using a physical breast exam simulator [100]. After the physical exam, the MP discussed her findings and next steps with Amanda. At several points throughout the interaction, Amanda challenged the MP by asking difficult questions or making troubling statements, e.g., "Do you think this could be cancer?" or "I don't know, do I have to get a mammogram?" The purpose of these challenges

is to provide the MP with opportunities to practice these difficult moments without any negative consequences (as there would with a real patient if the MP responds inappropriately).

The MP's verbal and nonverbal interaction with Amanda is facilitated using sensors. Microphones and speech recognition capture the MP's speech. The text of the speech is used to search a database of responses that Amanda can vocally and gesturally respond with. When an appropriate response is found, the corresponding audio is played with appropriate lip-synching. The MP's head is tracked, allowing Amanda to look at the MP. The MP can touch Amanda through a physical correlate, a mannequin outfitted with force sensors and the breast simulator. All verbal and nonverbal interaction was logged for later aggregate visualization in IPSViz^N.

The breast-cancer scenario provides opportunities for medical professionals to demonstrate a variety of medical communication skills, including completeness and organization of discussion, empathy, rapport-building, and reactions under pressure. These opportunities are also logged by the system so that they can be reviewed in IPSViz^N.

6.4 Preparing for Visualization

Before IPSViz^N can generate visualizations, the interactions with the virtual humans must be captured, filtered and processed. To capture, filter, and process, IPSViz^N uses the same visualization pipeline as that used by IPSViz. This section summarizes the stages of the pipeline. More details on these stages can be found in Chapter 4, Sections 4.3 and 4.4.

6.4.1 Capture

In the capture stage, the interaction between the user and virtual human is captured through a logging facility. All logged interaction data is timestamped and synchronized temporally. The logging facility captures both the user and the virtual human's verbal and nonverbal behavior. One or more video streams (including audio) of the interaction are captured. The video streams are recorded from perspectives that approximate the perspective of the virtual human. 6-DOF tracking data, such as user and virtual

human head pose and orientation, are also logged. Lastly, a transcript of virtual human speech and user speech, as determined by speech recognition software (Dragon Naturally Speaking 9.5), is logged. Together, the captured data provides a densely sampled (in time) set of *interaction signals* which characterize the interaction between the user and virtual human.

6.4.2 Filter and Process

The filter and process stage transforms captured interaction signals into whatever form is needed for visualization. Filtering 1) removes noise from the interaction signals (e.g., jitter in tracking data) and 2) simplifies interaction signals (e.g., by removing or aggregating samples) to highlight overall trends in and enable more rapid processing and visualization of the signals. Processing refers to computing new signals from existing signals (e.g., computing mutual gaze using head positions and orientations of the user and virtual human). Once the interaction signals are filtered and processed, they are passed on to the visualization stage, where users review visualizations to identify and compare trends and outliers in the data.

6.5 Identifying and Comparing Trends and Outliers with Visualization

As described previously, traditional approaches to identifying and comparing groups trends and outliers require 1) linear, one-at-a-time review of each interaction, 2) analyzing subtleties and relationships in each interaction's verbal and nonverbal behavior, and 3) linearly tracking and aggregating the subtleties and relationships to form judgments about the data's trends and outliers. Unfortunately, these tasks are time-consuming and cognitively difficult for users. IPSViz^N uses information visualization techniques to help users 1) review interactions non-linearly, 2) highlight subtleties and relationships in verbal and nonverbal behavior, and 3) visually aggregate and track them so that users require little cognitive effort to identify and compare trends and outliers.

6.5.1 Overviews of Verbal Behavior

IPSViz^N visually aggregates verbal communication in H-VH interactions using overview visualizations of verbal behavior. Verbal behavior is depicted visually by laying out a set of dots on a 2D grid. Each dot represents a verbal utterance by the virtual human to a user and is color-coded by the utterance's topic. A topic in this context is a theme, categorization, or meta-summary of an utterance. For example, the utterance "Hello, how are you?" might be labeled with the topic "greeting." Encoding utterances with colors allows users to rapidly interpret verbal communication in the H-VH interactions without having to read entire transcripts or watch videos of all the interactions.

The spatial organization or layout of the topic dots visually aggregates the data such that users require little cognitive effort to identify and compare trends and outliers in verbal behavior. Three types of layouts are available in IPSViz^N, each of which allows users to identify different relationships in the interactions' verbal communication:

- The **interaction timeline** (Figure 6-1) lays out topic dots by interaction (vertical axis) and time (horizontal axis). Each horizontal line of the plot represents the topic use in a specific interaction with a virtual human. The vertical alignment of each interaction's topic use allows identification of trends and outliers in interaction thoroughness, interaction organization, and length of interaction, as well as comparison of topic use in different interactions.

- The **topic timeline** (Figure 6-2) lays out topic dots by topic (vertical axis) and time (horizontal axis). Each horizontal line of the plot represents the use of a specific topic across all H-VH interactions. This plot shows gross trends and outliers in topic use (thoroughness and organization), independent of individual interactions.

- The **topic histogram** (Figure 6-3 lays out topic dots by frequency count (vertical axis) and topic (horizontal axis). The height of each vertical line of the plot indicates the relative frequency of that topic's use among the H-VH interactions, allowing quick identification of both dominant and rarely discussed topics.

6.5.2 Supporting Discovery and Comparison with Interaction

To help users review the data non-linearly and highlight subtleties and relationships in verbal and nonverbal behavior, four types of interaction are provided to the user: filtering, rescaling, querying, and semantic zooming.

6.5.2.1 Filtering to reveal trends and outliers

To reveal group and topic-use trends and outliers, IPSViz^N has interactive tools for filtering topic dots by topic and group. Trends and outliers in group and individual topic use can be occluded by overlapping topic dots. Thus, filtering by topic and group removes the dots to reveal hidden information. In addition, displaying too many topic dots on one visualization can be cognitively overwhelming to the user. Filtering by topic-use and group reduces the amount of data on the screen so that users can analyze the data without being overwhelmed. For example, questions like "Did group X discuss topic Y with the virtual human early in the interview?" or "Are groups X and Z different on discussion of topic Y?" are easier to answer if only the data associated with topic Y and groups X and Z are visible. Figure 6-4 provides an example of filtering by topic and subgroup.

The user interface for filtering consists of two menus, a legend of topics and a list of subgroups within the data. The topic legend, displayed just above the plots, both defines the mapping of colored dots to topics and also serves as a filtering menu. Selecting a topic name from the legend fades dots not corresponding to the selected topic out of view. Likewise, the interaction group list, displayed to the right of the plots, both defines the subgroups in the dataset and also serves as a filtering menu. Selecting a group name from the list fades dots not corresponding to the selected group out of view.

6.5.2.2 Rescaling time for identification and comparison

The scale of the time axis affords identification and comparison of different types of trends and outliers.

- With the **default scale**, the plots display each topic dot on their normal time scale (minutes and seconds from the beginning of the dot's corresponding interview). This time scale allows easy identification of time differences between interactions. For

example, in the plot displayed in Figure 6-1, one can see that the shortest interaction took about 3.5 minutes and the longest about 18 minutes.

- Alternatively, the **normalized scale** displays each interaction on a time scale between 0 and 1. Normalizing the time axis helps the user identify and compare order and frequency of topic discussion among a set of interactions, independent of actual length of each interaction. It would be difficult to analyze order and frequency of topic discussion otherwise, since interactions with smaller time lengths would be compressed into a very small area of the plot. Normalizing the time (Figure 6-5) scale addresses this problem by scaling each interaction's time dimension such that all interactions appear to have equivalent time lengths. This effectively stretches shorter interactions spatially to enable easier identification and comparison of topic discussion among all interactions.

6.5.2.3 Identifying outliers with selection queries

IPSViz^N uses query by visual selection to allow quick identification of outliers in the data. Query by visual selection involves clicking and dragging the mouse over a set of topic dots. The query returns the set of interactions associated with the selected dots. Thus, query by selection helps users map dots back to individual interactions. This is useful for identifying outliers because users can visually find an unusual pattern in the data (a potential outlier), and then identify the interactions associated with each pattern by selection.

Figure 6-6 demonstrates a medical-domain-oriented use of selection queries. An educator queries a set of history of present illness (HPI) dots near the end of the interviews. In the medical domain, the dots represent outliers because medical professionals should discuss HPI with the patient earlier. Discussing HPI later in the interview is a sign of conducting a disorganized medical interview. The query returns the list of potentially disorganized users to the educator, and the educator can use this list to investigate the potentially disorganized users further.

6.5.2.4 Semantic zooming to review specific moments

While the visualizations provide visual summaries of the interactions, some types of analysis require drilling down to the details of a specific moment in a specific interaction. Details not provided by the summary visualizations include the tone of voice of the user

when responding to a difficult question from the virtual human or the way in which a user transitions from one topic to the next. Semantic zooming [138], changing the representation of the data as it fills more screen space, is applied here to help the user review specific moments in the interaction.

To enable targeted review of specific moments, video is associated with each topic dot (Figure 6-7). As the video and topic data is synchronized in time during capture, each topic dot is linked to its corresponding moment in its associated video. Double-clicking on a dot semantically zooms into the dot and displays a video window. Video playback starts at the selected moment. Clicking again stops playback and hides the video window. Thus, the user can easily select and review specific moments of interest for detailed review.

6.6 Preliminary Evaluation

Fifteen participants informally evaluated IPSViz^N. The goals of the evaluation sessions were to 1) determine if participants could discover trends and outliers in a set of H-VH interactions using IPSViz^N, 2) learn the extent to which errors in the data affect observations negatively, and 3) elicit feedback on how to improve IPSViz^N.

The participants were computer scientists (9), medical educators (4), and a medical student. They either watched a presentation and live demo of IPSViz^N or were sent a web link to try IPSViz^N on their own time.

6.6.1 Dataset

For the evaluation, IPSViz^N displayed a set of 29 medical professionals (MPs) interacting with Amanda, the VH breast cancer patient described previously. Each MP conducted one interaction with Amanda for a total of 29 interactions available for review in IPSViz^N. The MPs were composed of 23 males and six females. Sixteen were medical students, three were interns, five were residents, and five were experienced clinicians. Participants could explore and compare the topic discussion of these different subgroups using the visualizations described previously. The topics displayed in the dataset were chief complaint, history of present illness (hpi), review of systems (ros), medical history,

social history, family history, and next steps (mammogram). These topics represent the critical sets of information the user should gather from a breast cancer patient. These topics should be discussed with the patient in this order. In addition, patient challenges (e.g., "Could this be Cancer?") and corresponding empathy from users are also marked in the visualizations. Thus, filtering (by subgroup or topic), reviewing the topic dots, and watching their associated videos enables comparing and evaluating how users respond to challenges and use of empathy (if any).

6.6.2 Questionnaire

A domain-oriented questionnaire was placed on the same web page as IPSViz^N for the medical educators and student to fill out. The questionnaire was designed to encourage review and discovery of group medical communication skills as well as solicit domain-oriented feedback. The review and discovery portion of the questionnaire asked:

1. In what order did users tend to conduct their interview? Does any subgroup tend to conduct their interview in a different order than the entire group?
2. What are the differences between students and clinicians? Is one group more organized? Is one group more thorough and complete?
3. What can you identify with respect to individual students (e.g., best student? worst student? student who spent the most time? etc.)
4. What can you identify with respect to subgroups?
5. Use the plotter freely. If you make any new observations, please write them here.

The feedback portion of the questionnaire asked:

1. About how long did it take you to answer the questions above?
2. Which questions were difficult to answer? Which took the longest to answer? Why?
3. Was there anything you wanted to do that you couldn't do?
4. How would you improve the topic plotter?
5. If you have any other feedback, please write it here.

6.6.3 Results

Participants identified trends in the dataset, and compared trends among groups. Several participants felt the small number of participants in certain subgroups made

it difficult to identify trends with certainty for those subgroups. Trust in the interface was high because the visualizations fit participant models of what a medical interview looks like for the representative groups. Computer scientists identified outliers, but medically-oriented participants did not, implying a different interface may be needed to highlight outliers to the medical audience. Finally, participants provided a number of ideas on how IPSViz^N should evolve as a research and education tool.

6.6.3.1 Trends

IPSViz^N allowed identification of trends in the overall interactions. The identified order of topic discussion for the entire group of interactions was consistent across all participants. In addition, participants indicated not enough empathy was used in the interactions. Identifying trends in subgroups was more difficult. Participants indicated they did not feel comfortable providing trends with smaller sample sizes (e.g., there were only six women in the group).

Participants compared trends across groups. They noted that clinicians discussed family history less and focused mostly on the history of present illness. Students, on the other hand, stuck to the standard organization of the medical interview, proceeding from topic to topic in the specified order. Some participants expected this difference between clinicians and students but did not have any prior evidence to confirm it. Clinicians, they told us, have the experience to know exactly what questions to ask the patient, whereas inexperienced students need to conduct organized interviews to make sure they get all the information needed to diagnose the patient. Thus, IPSViz^N allowed participants to confirm a hypothesis about group trends, and participants indicated this trend with the virtual human reflects a trend with humans as well.

Review with IPSViz^N encouraged generation and testing of new hypotheses about the interactions as well. One participant hypothesized a relationship between discussion of social history and use of empathy. After using IPSViz^N, the participant determined no such relationship exists.

6.6.3.2 Outliers

Outliers were more difficult to identify. The survey encouraged participants to look for outliers by asking them to identify the best and worst MPs in the group. However, participants indicated this was impossible without incorporating other metrics into the interface, such as whether the MP was able to gather all the important information from the patient.

Although she did not discuss any one specific individual, one participant noted IPVSiz^N could highlight concerns about an individual. With IPSViz^N, users could identify "those who never show empathy, those who are too brief, too long, [and] those who never mention mammograms (next steps). These users may not represent the best or worst MPs, but they might be forgetting one important detail in their interactions with patients. Furthermore, these mistakes could be a sign of a larger problem with a person's interaction with patients that should be studied further with other tools.

One reason outliers were difficult to find was that the interface did not highlight the tools for identifying outliers. One of the main tools for identifying outliers, the selection queries, required clicking-and-dragging. Although instructions were provided, unless the user tried it explicitly, they may not have known it even existed. Adding a selection query mode which is activated by a button may call more attention to this feature.

Additionally, outliers may not have been dramatic enough to stand out from the large trends in topic discussion. Replacing rows of topic dots that represent trends with a single cluster dot might allow non-trend data (outliers) to stand out more from the dataset.

6.6.3.3 Errors in the data

One trend in participant comments is the need for more transparency in the interface. Participants complained that they did not know how dots were assigned topic labels, and thus, did not know to what extent they should trust it reflected the actual interpersonal behavior of the users. One participant asked if a clinician coded the data, and another wanted to know how the challenge and empathy topic dots were labeled.

The perceptions of the participants were correct; the topic dots can be mislabeled. To label the dots, the virtual human's utterances are pre-labeled with topics when the virtual human's script is created. Pre-tagging VH utterances is possible because every utterance spoken by the virtual human is predefined in the script. Later, during the H-VH interaction, the VH's utterances are logged with their pre-assigned topics for later visualization. This leads to mislabeled topic dots if the topic the virtual human discusses is not the same topic being discussed by the user. This can happen when the virtual human misunderstands the user's question, and thus responds with incorrect speech. Current estimates indicate this occurs, on average, 26% of the time with a standard deviation of 8.4%. Fortunately, much fewer dots are improperly labeled because incorrect responses are often of the same topic as correct responses.

Indeed, despite the errors in the topic labels, participants were able to identify trends in the data, and they appeared to trust these trends were real rather than a result of error. Participants trusted the trends because the trends fit their expected model of what happens in a patient-doctor interview. For example, the participants said clear trends in the discussion of history of present illness were the same as that expected in a similar interaction with a real patient. Thus, the information provided by IPSViz^N seemed trustworthy. Overall, the error rate affects user observations, but not enough to outweigh the benefits of the system.

Nonetheless, future versions should incorporate some transparency to the process of labeling the data. Adding video playback to the interface partially addresses this issue, as users can watch video associated with a dot to see if the dot is labeled with the correct topic. Historical trends in the error rates could also be displayed in the interface. In the long term, relabeling tools could be added to the interface so that users could relabel topic dots when video review reveals they are labeled incorrectly.

6.6.3.4 Suggested improvements

The medically-oriented participants had a number of suggestions for IPSViz^N which mirrored suggestions provided by the computer scientists.

- One participant wanted to compare groups on a specific topic. The topic filters address this in part, but comparisons would be easier if the topic discussion of each group were placed side-by-side in the interface. The current approach of only displaying the selected group's topic dots makes it harder to do this kind of comparison.

- Another participant spent 40 minutes trying to identify trends in topic flow for individual groups. While side-by-side comparisons would help here, clustering or summarizing groups into a single topic flow would make trend comparison significantly faster. Then additional tools could be provided to drill down into the interactions of specific individuals.

- Multiple participants wanted more metrics built into IPSViz^N, including nonverbal behavior, whether the MP collected all the information needed, and the MP's performance on the physical exam.

- Medical educators routinely evaluate students using simple statistical summaries of student performances. Medical educators noted these were missing in IPSViz^N and asked for averages, standard deviations, percentages, and other statistics to be displayed.

- One medical educator indicated it would be useful to have gold-standard interactions (e.g., an interaction with ideal topic flow) built into the interface for comparison to the groups. This would help students see how they specifically differ from the group and target areas to practice more.

- As outliers were difficult to identify, some computer scientists argued for having separate visualizations for trends and outliers, respectively.

- Lastly, multiple participants from both the medical and computer science domain asked for a tutorial or established procedure to follow when using ISPViz^N. This would help them learn how to use all the tools as well as provide some added consistency and confidence to evaluations with IPSViz^N.,

6.7 Conclusions and Future Work

This chapter presents IPSViz^N, a proof-of-concept visualization system for rapidly uncovering trends and outliers in groups of human-virtual human interactions. Trends and outliers are uncovered by reviewing aggregate visualizations of the interactions.

The visualizations summarize conversations between humans and virtual humans. Filtering (topics and groups), semantic zooming (video), time rescaling (normalization), and querying operators facilitate rapidly uncovering trends and outliers in verbal and nonverbal behavior from a large data set of H-VH interactions.

An evaluation of IPSViz^N with representative end-users highlighted several use cases for analyzing groups of H-VH interactions. These include identifying unusual or interesting cases within a group (outliers), identifying common cases within a group (trends), comparing trends and outliers between individuals and/or subgroups, and generating and confirming hypotheses about H-VH interactions.

The evaluation also shows one can generate visualizations of recorded H-VH interactions that provide insights that would be difficult or impossible to learn from real-world interactions. End-users were able to rapidly (within minutes) identify trends in overall group interpersonal skills, including verbal behavior, organization, completeness, empathy, and communicating under stress. Identifying these trends without IPSViz^N would have required hours of manual effort, including video review, transcription, and meta-analysis of transcripts (topic labeling).

Lastly, the evaluation indicates one can evaluate a group's real-world interpersonal skills by evaluating that group's interactions with virtual humans. End-users identified trends in interpersonal behavior with a virtual human, and compared trends among subgroups of interactions (e.g., clinicians vs. students). Some observed trends fit end-user expectations of interpersonal behavior in a similar real-world interaction, providing strong evidence that IPSViz^N displays real-world interpersonal skills to end-users.

The next step is to integrate virtual human patient experiences and IPSViz^N into a medical communication curriculum. This will allow studying how end-users (medical educators, students, and VH researchers) use IPSViz^N in a more realistic scenario where much larger groups of interactions will need to be analyzed. In this scenario, N will grow exponentially as students perform more and more interactions with VH patients over the

course of a semester. The dynamics of a curriculum will require new interaction modes such as tracking user progress over time, and searching for outliers in very large datasets.

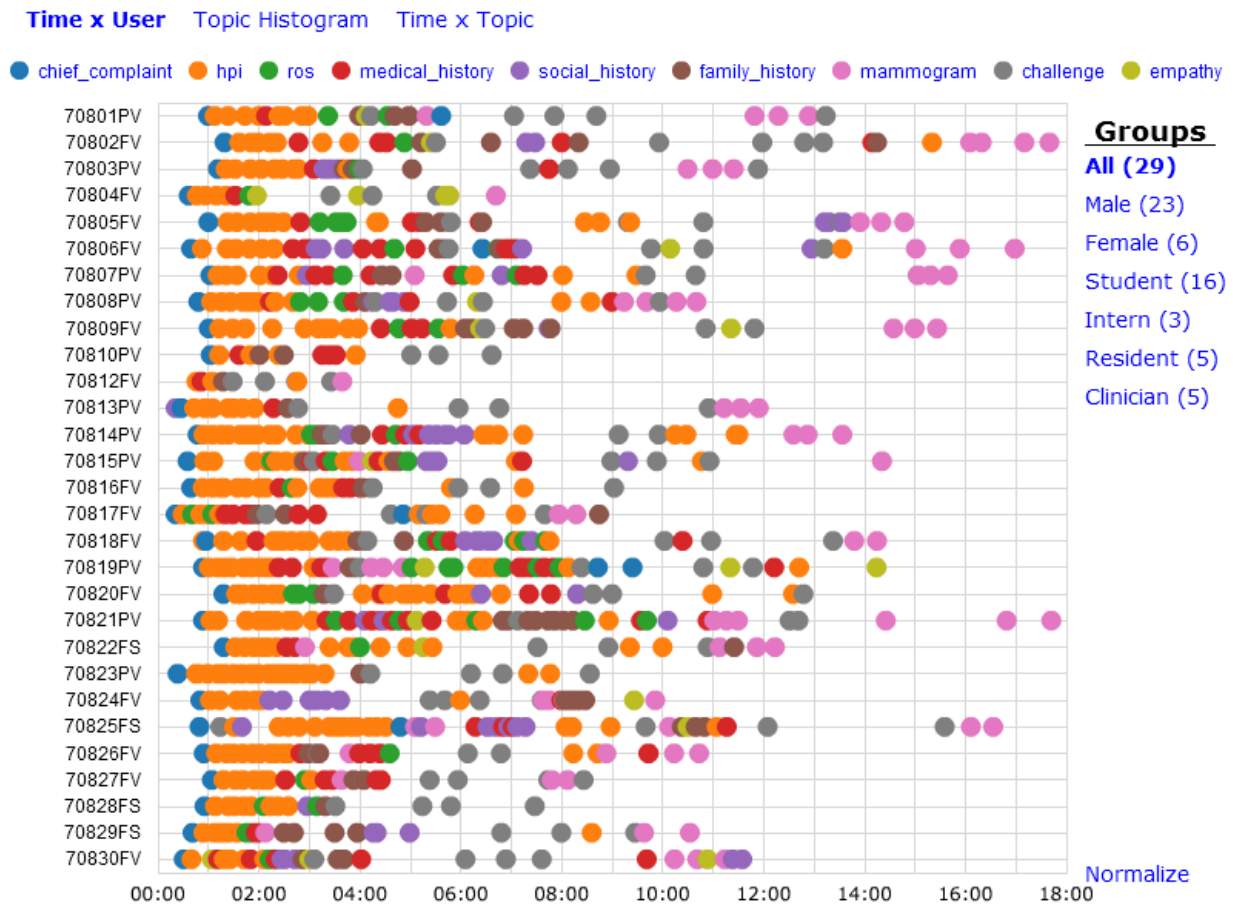


Figure 6-1. Interaction timeline shows use of topics over time with respect to each interaction.

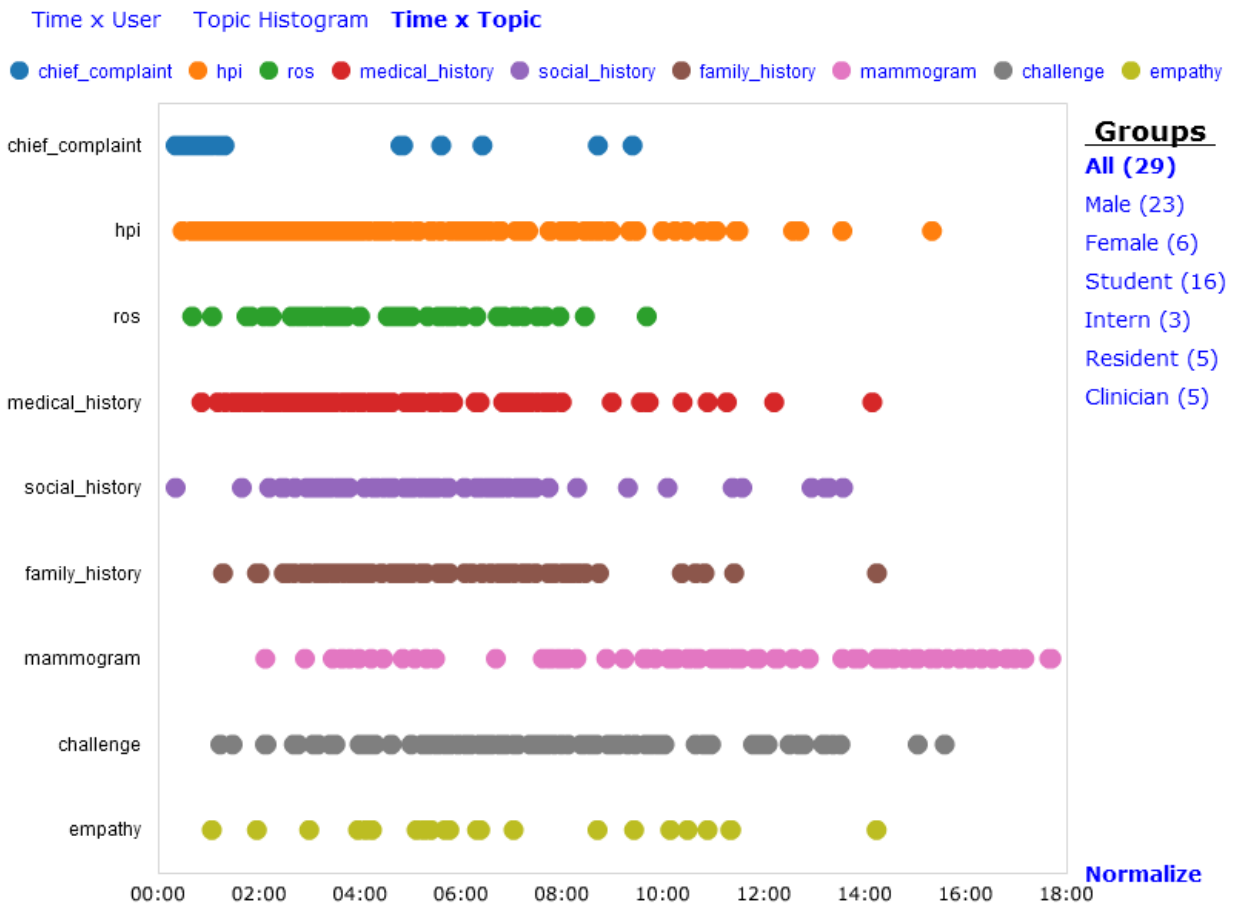


Figure 6-2. Topic timeline shows use of topics over time for all users.

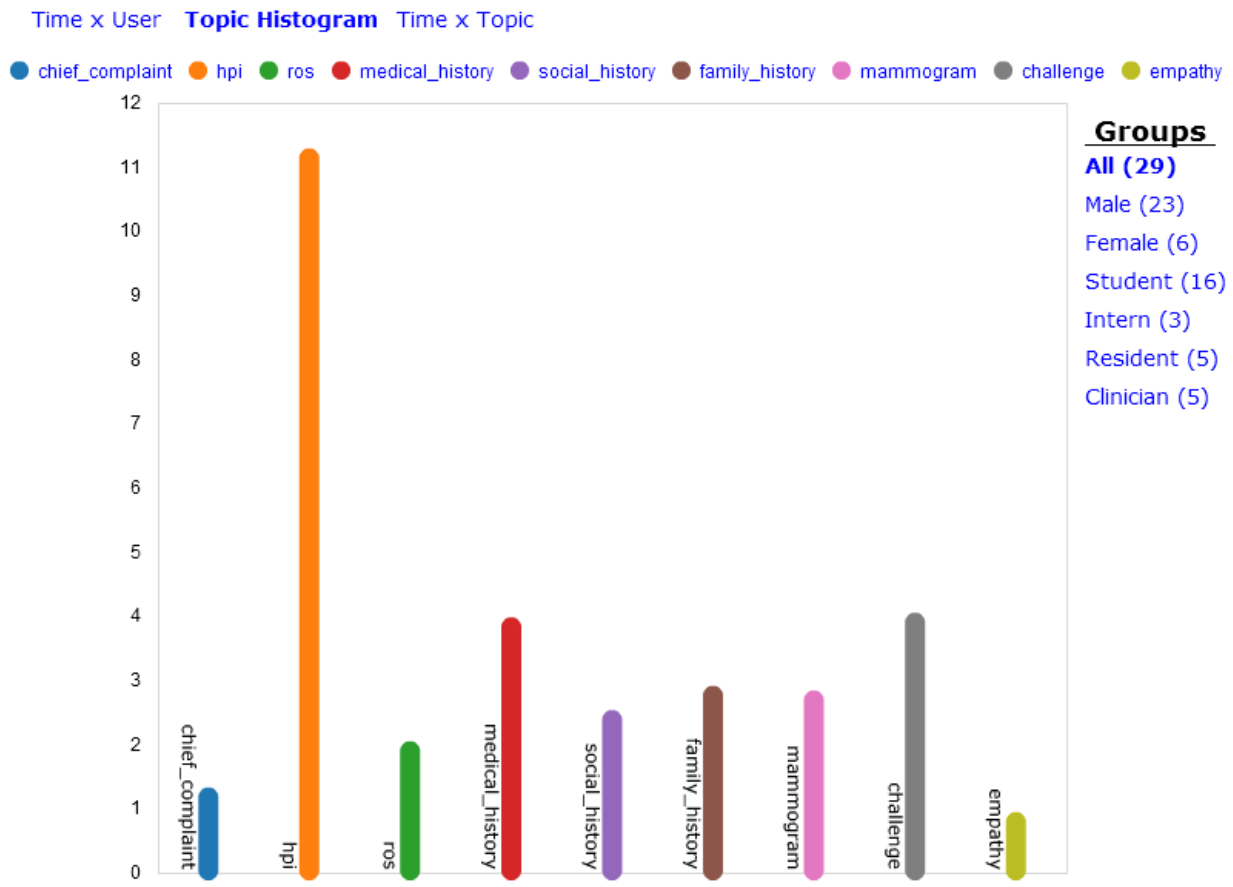


Figure 6-3. Topic histogram shows frequency of topic use for a selected group of users.

Figure 6-4. Interaction timeline visualization (Figure 6-1 with a series of filters applied): A) Discussion of history of present illness (hpi) only. B) Reviewing hpi, but further filtering to only review the clinicians subgroup. C) Reviewing hpi only for the students subgroup.

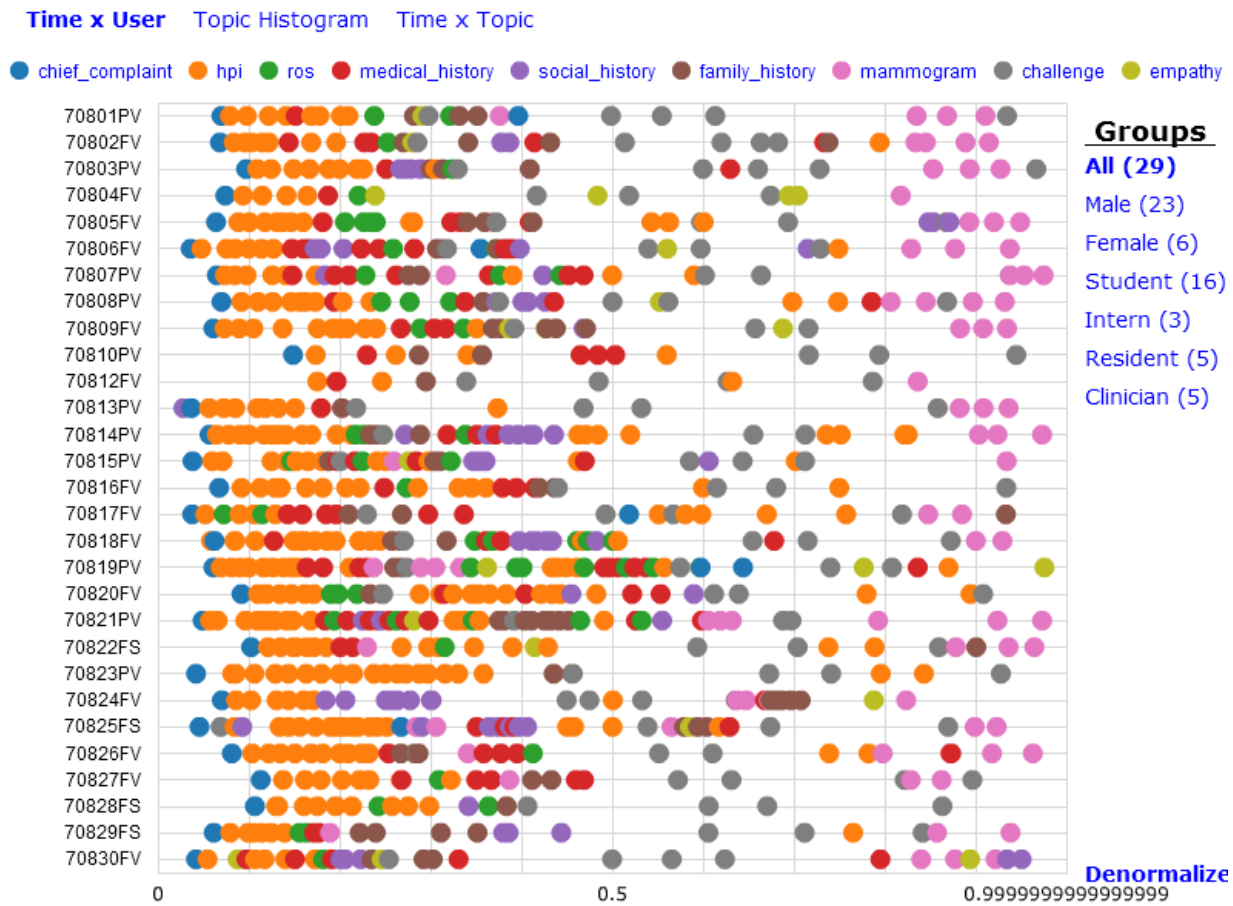


Figure 6-5. Use of topics on a normalized timeline

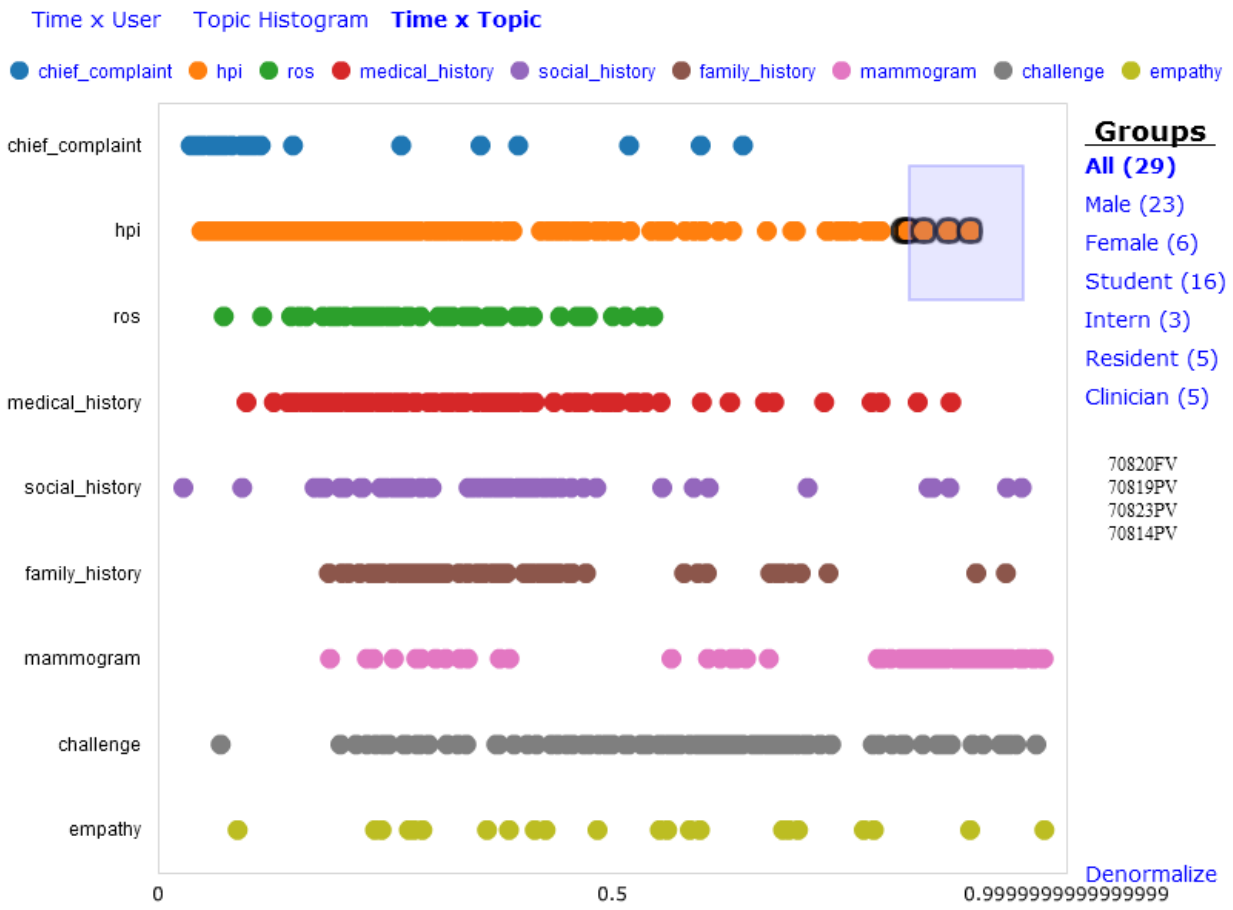
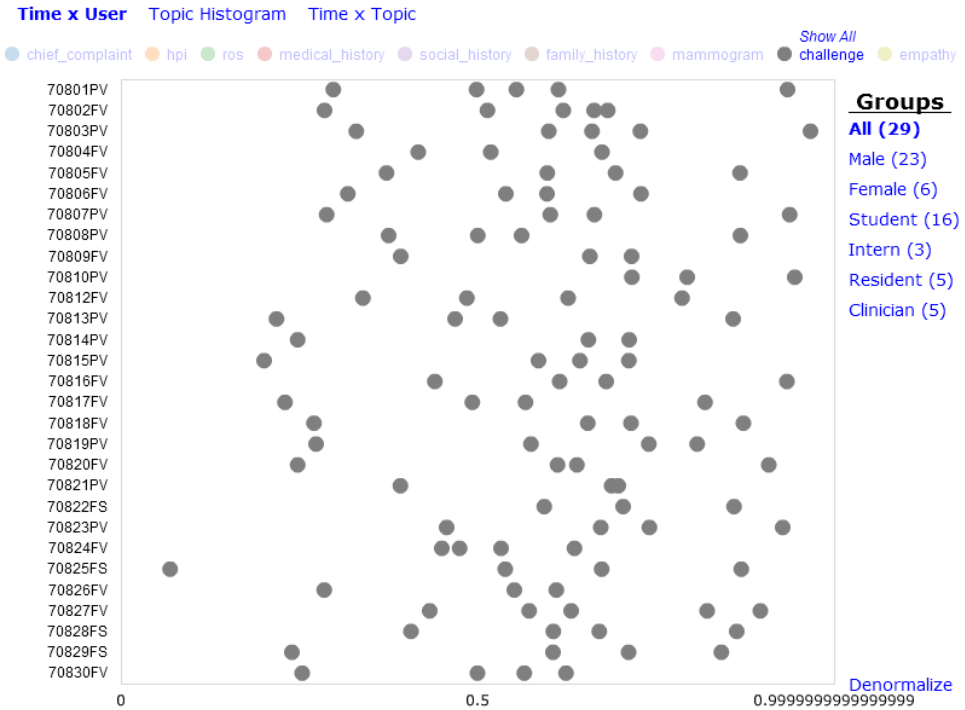


Figure 6-6. Querying timespans by selection



A

B

Figure 6-7. Reviewing video of patient challenges. A) Interaction timeline filtered by patient challenges. B) Upon selecting challenge, corresponding video appears for review.

CHAPTER 7
SUMMARY AND FUTURE WORK

We proposed after-action reviews (AARs) for human-virtual human (H-VH) interactions. Three proof-of-concept AAR systems were presented: the Interpersonal Scenario Visualizer (IPSViz), the VSP system, and IPSViz^N. These three systems use information visualization and immersive virtual reality techniques to provide insight into domain-oriented interpersonal skills. A series of user studies found that reviewing H-VH interactions with IPSViz, the VSP system, and IPSViz^N elicits reflection on, changes perceptions of, and motivates change in domain-specific interpersonal skills.

7.1 Lessons from IPSViz

Reviewing one's interaction with a virtual human using IPSViz elicited reflection on and discussion of self-directed change of one's interpersonal skills. Skills reflected on included verbal and nonverbal behavior, building rapport, and communicating effectively under stress. The most surprising result was that some participants claimed they would change interpersonal behavior in future interactions with real humans as a result of the AAR. This means that 1) these participants believed that their interpersonal skills with a virtual human reflects their interpersonal skills with real humans, and 2) that after-action reviews of one's interaction with a virtual human can motivate changing one's interactions with real humans.

7.2 Lessons from the VSP System

Reviewing an interaction with a virtual human by reliving it as the virtual human elicits social perspective-taking - reflection on the experience from the virtual human's perspective - and self-directed change in interpersonal skills. By reliving their interaction with a VH from the VH's perspective, participants were forced to reflect on their behavior from the perspective of a virtual human. This reflection improved participant empathy in a second interaction with another virtual human, indicating that VSP elicits social perspective-taking and self-directed improvements in behavior. The experience was

particularly powerful because it enabled people to relive - and better understand - the unfamiliar experiences of another person.

7.3 Lessons from IPSViz^N

Reviewing groups of human-virtual human interactions using aggregate visualizations encouraged reflection on the real-world interpersonal skills of groups. Participants hypothesized and identified group trends and outliers in the interactions, and compared group trends to determine if groups differ on specific interpersonal skills, including verbal behavior, organization, completeness, and empathy. Participants noted the tool would be useful for identifying H-VH users who need to improve specific aspects of their interpersonal skills. Thus, IPSViz^N has the potential to motivate self-directed change.

By proposing after-action reviews for human-virtual human experiences, this dissertation expands the capabilities of human-virtual human experiences. Adding after-action reviews to human-virtual human experiences elicits reflection on, changes perceptions of, and motivates change in domain-specific interpersonal skills.

7.4 Future Work

The ultimate goal of interpersonal simulators, like the doctor-patient system used in this dissertation, is to improve interactions with real humans. This dissertation stops just short of that goal in finding that, after conducting an AAR, users changed behavior with a virtual human, and said they would change behavior with real humans. The next step is to test if AARs of H-VH experiences can lead to measurable change in one's interactions with real humans.

Also important is determining if the AARs change user behavior in desired ways. This dissertation shows that AARs impact people, but it is not clear if the AARs impact people to meet desired goals. The next step is to explore the parameter-space of AARs to identify how to impact specific skills through AAR of H-VH interactions. One possible approach would be to incorporate expert knowledge into the AAR system.

It would be useful to compare existing approaches to after-action review of interpersonal interactions and the approaches discussed in this dissertation. The approaches described here incorporate previous approaches (e.g., video review). Thus, we expect the approaches described here will provide additional benefits on top of existing approaches. Nevertheless, a comparison would highlight exactly what those additional benefits are and when to choose between simpler approaches and those proposed here.

Lastly, the success of visualization and immersive virtual reality techniques in AARs of human-virtual human interactions motivates their use for AARs of human-human interactions. It is significantly easier to create such AARs for H-VH interactions because much of the interaction (the virtual human and the environment) is computer generated. With H-H interactions, all aspects of the interaction are real, significantly increasing the challenges of capture and processing. If these challenges can be overcome, then we would be able to provide people with tools to review, analyze, and learn from their everyday lives.

REFERENCES

[1] W. Acevedo and P. Masuoka. Time-series animation techniques for visualizing urban growth. *Computers and Geosciences*, 23(4):423–435, 1997.

[2] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented dataA systematic view. *Computers & Graphics*, 31(3):401–409, 2007.

[3] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual Methods for Analyzing Time-Oriented Data. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, pages 47–60, 2008.

[4] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. PlanningLines: Novel Glyphs for Representing Temporal Uncertainties and their Evaluation. In *The Ninth International Conference on Information Visualization*, pages 457–463, 2005.

[5] F. S. Alallah, M. Nezhadasl, P. Irani, and D. Jin. Visualizing the decision-making process in a face-to-face meeting. In *11th International Conference Information Visualization (IV'07)*, pages 168–176, 2007.

[6] N. E. Alessi and M. P. Huang. Evolution of the virtual human: From term to potential application in psychiatry. *CyberPsychology & Behavior*, 3(3):321–326, 2000.

[7] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*, 14(6):503–541, 2003.

[8] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, UK, 1976.

[9] J. Astington and D. Olson. Metacognitive and Metalinguistic Language: Learning to Talk about Thought. *Applied Psychology*, 39(1):77–87, 1990.

[10] S. Babu, T. Barnes, and L. Hodges. Can immersive virtual humans teach social conversational protocols? In *IEEE Virtual Reality 2007*, pages 215–218, Charlotte, NC, March 2007.

[11] K. Bachour, F. Kaplan, and P. Dillenbourg. Reflect: An interactive table for regulating face-to-face collaborative learning. In *Proceedings of the European Conference on Technology Enhanced Learning 2008*, 2008.

[12] N. I. Badler, C. A. Erignac, and Y. Liu. Virtual humans for validating maintenance procedures. *Comm. of the ACM*, 45(7):56 – 63, 2002.

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

[14] J. N. Bailenson, E. Aharoni, A. C. Beall, R. E. Guadagno, A. Dimov, and J. Blascovich. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *The 7th Annual International Workshop on Presence*, Valencia, Spain, 2004.

[15] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *PRESENCE: Teleoperators and Virtual Environments*, 10(6):583–598, 2001.

[16] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoper. Virtual Environ.*, 14(4):379–393, 2005.

[17] L. Baird, J. Henderson, and S. Watts. Learning from Action: An Analysis of the Center for Army Lessons Learned (CALL). *Human Resource Management*, 36(4):385–395, 1997.

[18] L. Baird, P. Holland, and S. Deacon. Learning from action: Embedding more learning into the performance fast enough to make a difference. *Organizational Dynamics*, 27(4):19–32, 1999.

[19] S. Balcisoy, R. Torre, M. Ponder, P. Fua, and D. Thalmann. Augmented reality for real and virtual humans. In *Computer Graphics International*, pages 303–308, 2000.

[20] D. Bauer and J. Hollan. IRYS: a visualization tool for temporal analysis of multimodal interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 285–288, 2003.

[21] B. Bederson, A. Clamage, M. Czerwinski, and G. Robertson. DateLens: A fisheye calendar interface for PDAs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(1):90–119, 2004.

[22] T. Bergstrom and K. Karahalios. Seeing More: Visualizing Audio Cues. *Proceedings of INTERACT 2007*, 2007.

[23] T. Bergstrom and K. Karahalios. Conversation clusters: human-computer dialog for topic extraction. In *CHI '08 Extended Abstracts on Human factors in Computing Systems*, pages 2829–2834, 2008.

[24] M. Billger. Colour appearance in virtual reality: A comparison between a full-scale room and a virtual reality simulation. In *The 9th Congress of the International Colour Association*, 2001.

[25] J. Blascovich. A theoretical model of social influence for increasing the utility of collaborative virtual environments. In *CVE '02: Proceedings of the 4th International Conference on Collaborative Virtual Environments*, pages 25–30, New York, NY, USA, 2002. ACM.

[26] A. Boucouvalas. Real Time Text-to-Emotion Engine for Expressive Internet Communications. *Emerging Communication*, 5:305–320, 2003.

[27] E. Boyd and A. Fales. Reflective Learning: Key to Learning from Experience. *Journal of Humanistic Psychology*, 23(2):99, 1983.

[28] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proc. SPIE*, volume 5669, pages 175–186, 2005.

[29] L. Byron and M. Wattenberg. Stacked Graphs–Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[30] J. V. Carlis and J. A. Konstan. Interactive visualization of serial periodic data. In *UIST '98: Proceedings of the 11th annual ACM symposium on User interface software and technology*, pages 29–38, New York, NY, USA, 1998. ACM.

[31] S. Carroll, S. Beyerlein, M. Ford, and D. Apple. The Learning Assessment Journal as a tool for structured reflectionin process education. In *Proceedings of Frontiers in Education Conference (FIE'96)*, volume 1, 1996.

[32] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 114–123. Association for Computational Linguistics Morristown, NJ, USA, 2001.

[33] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[34] G. Collier. *Emotional Expression*. Lawrence Erlbaum Associates, London, 1985.

[35] J. L. Coulehan and M. R. Block. *The Medical Interview: Mastering Skills for Clinical Practice*. F.A. Davis Co., Philadelphia, 3rd edition, 1997.

[36] S. Craig, B. Gholson, and D. Driscoll. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*, 94(2):428–434, 2002.

[37] C. Daassi, L. Nigay, M. Fauvet, L. CLIPS-IMAG, and F. Grenoble. A taxonomy of temporal data visualization techniques. *Information-Interaction-Intelligence*, 5(2):41, 2006.

[38] M. Darling and C. Parry. After-Action Reviews: Linking Reflection and Planning in a Learning Practice. *Reflections: The SoL Journal*, 3(2):64–72, 2001.

[39] D. Davis, P. Mazmanian, M. Fordis, R. Van Harrison, K. Thorpe, and L. Perrier. Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review. *JAMA*, 296(9):1094, 2006.

[40] E. Deaton, C. Barba, T. Santarelli, L. Rosenzweig, V. Souders, C. McCollum, J. Seip, W. Knerr, and J. Singer. Virtual environment cultural training for operational readiness (vector). *Virtual Reality*, 8(3):156–167, 2005.

[41] Department of the Army, Washington, DC. *Training Circular 25-20: A Leader's Guide To After-Action Reviews*, September 1993.

[42] M. Deutsch and E. National Association for Mediation in. Educating for a peaceful world. *American Psychologist*, 48:510–510, 1993.

[43] R. Dickerson, K. Johnsen, A. Raij, B. Lok, T. Bernard, A. Stevens, and D. S. Lind. Virtual patients: Assessment of synthesized versus recorded speech. In *Medicine Meets Virtual Reality 14*, Long Beach, CA, 2006.

[44] R. Dickerson, K. Johnsen, A. Raij, B. Lok, J. Hernandez, and A. Stevens. Evaluating a script-based approach for simulating patient-doctor interaction. In *The International Conference on Human-Computer Interface Advances for Modeling and Simulation*, New Orleans, LA, 2005.

[45] S. DiGangi. Self-Graphing of On-Task Behavior: Enhancing the Reactive Effects of Self-Monitoring on On-Task Behavior and Academic Performance. *Learning Disability Quarterly*, 14(3):221–30, 1991.

[46] J. DiMicco, K. Hollenbach, A. Pandolfo, and W. Bender. The Impact of Increased Awareness While Face-to-Face. *Human-Computer Interaction*, 22(1-2):47–96, 2007.

[47] J. Dovidio, K. Kawakami, and S. Gaertner. Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1):62–68, 2002.

[48] F. Duffy and E. Holmboe. Self-assessment in Lifelong Learning and Improving Performance in Practice: Physician Know Thyself. *JAMA*, 296(9):1137, 2006.

[49] E. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, SK Card. Visualizing the Evolution of Web Ecologies. In *Conference on Human Factors in Computing Systems (CHI98)*, pages 400–407, 1998.

[50] P. Eisert, J. Rurainsky, and P. Fechteler. Virtual mirror: Real-time tracking of shoes in augmented reality environments. *IEEE International Conference on Image Processing 2007*, 2:557–560, 2007.

[51] P. Emmelkamp, M. Krijn, A. Hulsbosch, S. de Vries, M. Schuemie, and C. van der Mast. Virtual reality treatment versus exposure in vivo: a comparative evaluation in acrophobia. *Behaviour Research and Therapy*, 40:509–516, 2002.

[52] T. Erickson, D. Smith, W. Kellogg, M. Laff, J. Richards, and E. Bradner. Socially Translucent Systems: Social Proxies, persistent conversation, and the design of babble. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: The CHI is the Limit*, pages 72–79, 1999.

[53] J. Fails, A. Karlson, L. Shahamat, and B. Shneiderman. A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events over Time. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST'06)*, 2006.

[54] Fifth Dimension Technologies. *Flight Visualization System*, August 2008. Fifth Dimension Technologies.

[55] J. Flavell. Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist*, 34(10):906–11, 1979.

[56] R. Ford, V. Lamar, R. Giuli, and S. Oberg. The Joint Training Experimentation Program Approach to a Distributed After Action Review (DAAR). *04E-SIW-063, Simulation Interoperability Workshop*, 2004.

[57] R. Ford, J. Shockley, M. Beebe, M. Faust, G. Lucha, M. Johnson, and J. Bernatz. The Joint Training Experimentation Program: Lessons Learned from the First Demonstration. *Simulation Interoperability Workshop*, 2003.

[58] Forterra. On-line interactive virtual environment (olive): A multi-user, virtual world software platform. Technical report, Forterra Systems, December 2006.

[59] G. Frank, C. Guinn, R. Hubal, P. Pope, M. Stanford, and D. Lamm-Weisel. Just-talk: An application of responsive virtual human technology. In *In Proc. of the Interservice/Industry Training, Simulation and Education Conference*, Orlando, FL, December 2002.

[60] D. Friedman, A. Brogni, A. Antley, C. Guger, A. Steed, and M. Slater. Standardizing data analysis in presence experiments. *Presence: Teleoperators and Virtual Environments*, 15(5):599–610, October 2006.

[61] G. Furnas. Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 16–23. ACM New York, NY, USA, 1986.

[62] G. Furnas. The FISHEYE View: A New Look at Structured Files. *Readings in Information Visualization: Using Vision to Think*, pages 312–330, 1999.

[63] A. Gande. *Instructor Station for virtual laparoscopic surgery.* PhD thesis, Masters thesis, University of Texas at Arlington, August 2003.

[64] M. Garau, M. Slater, S. Bee, and M. A. Sasse. The impact of eye gaze on communication using humanoid avatars. In *The SIGCHI Conference on Human Factors in Computing Systems*, Seattle, WA, 2001.

[65] M. Garau, M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proc. of the Conference on Human Factors in Computing Systems*, pages 529–536, 2003.

[66] D. Garvin. Building a learning organization. *Harv Bus Rev*, 71(4):78–91, 1993.

[67] R. Granlund. Web-based micro-world simulation for emergency management training. *Future Generation Computer Systems*, 17(5):561–572, 2001.

[68] J. Gratch and W. Mao. Automating After Action Review: Attributing Blame or Credit in Team Training. *Conference on Behavior Representation in Modeling and Simulation*, 2003.

[69] C. Greenhalgh, M. Flintham, J. Purbrick, and S. Benford. Applications of temporal links: Recording and replaying virtual environments. In *IEEE Virtual Reality 2002*, March 2002.

[70] D. Gresh, D. Rabenhorst, A. Shabo, and S. Slavin. Prima: A case study of using information visualization techniques for patient record analysis. In *Proceedings of IEEE Visualization*, 2002.

[71] M. R. Guibert, C. A. Pelz, E. N. Bachelder, and K. Barrera. Next generation visualization tool for mission planning, briefing, and after action review. In *IMAGE 2007 Conference*, Scottsdale, Arizona, July 2007.

[72] F. Gulliver. Post-project appraisals pay. *Harvard Business Review*, 65(2):128–32, 1987.

[73] A. Guye-Vuillème, T. Capin, S. Pandzic, N. Thalmann, and D. Thalmann. Nonverbal communication interface for collaborative virtual environments. *Virtual Reality*, 4(1):49–59, 1999.

[74] J. C. Hart. On recording virtual environments. In *IEEE Symposium on Research Frontiers in Virtual Reality*, page 8083, October 1993.

[75] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, pages 9–20, 2002.

[76] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2008.

[77] I. Heldal, R. Schroeder, A. Steed, A. S. Axelsson, M. Spant, and J. Widestrom. Immersiveness and symmetry in copresent scenarios. In *IEEE Virtual Reality 2005*, pages 171–178, 2005.

[78] M. Hickson III, D. W. Stacks, and N.-J. Moore. *Nonverbal Communication: Studies and Applications*. Roxbury Publishing Company, Los Angeles, CA, 2004.

[79] R. W. Hill Jr., J. Gratch, S. Marsella, J. Rickel, W. Swartout, and D. Traum. Virtual humans in the mission rehearsal exercise system. *Kynstliche Intelligenz (KI Journal)*, 17(4), 2003.

[80] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[81] M. Hojat, J. S. Gonnella, T. J. Nasca, S. Mangione, M. Vergare, and M. Magee. Physician empathy: Definition, components, measurement, and relationship to gender and specialty. *American Journal of Psychiatry*, 159(9):1563, 2002.

[82] J. Hong and J. Landay. WebQuilt: a framework for capturing and visualizing the web experience. In *Proceedings of the 10th international conference on World Wide Web*, pages 717–724. ACM New York, NY, USA, 2001.

[83] A. Hoover, E. Muth, and F. Switzer. Clemson university mout project. www.parl.clemson.edu/~ahoover/MOUT/. Clemson University, July 2008.

[84] D. Huynh, S. Drucker, P. Baudisch, and C. Wong. Time quilt: scaling up zoomable photo browsers for large, unstructured photo collections. In *Conference on Human Factors in Computing Systems*, pages 1937–1940. ACM New York, NY, USA, 2005.

[85] J. Jenvald and M. Morin. Simulation-Supported Live Training for Emergency Response in Hazardous Environments. *Simulation & Gaming*, 35(3):363, 2004.

[86] K. Johnsen. *Design and Validation of a Virtual Human System for Interpersonal Skills Education*. Dissertation, University of Florida, 2008.

[87] K. Johnsen, R. Dickerson, A. Raij, C. Harrison, B. Lok, A. Stevens, and D. S. Lind. Evolving an immersive medical communication skills trainer. *Presence: Teleoperators and Virtual Environments*, 15(1):33–46, February 2006.

[88] K. Johnsen, R. Dickerson, A. Raij, B. Lok, J. Jackson, M. Shin, J. Hernandez, A. Stevens, and D. S. Lind. Experiences in using immersive virtual characters to educate medical communication skills. In *IEEE Virtual Reality 2005*, Bonn, Germany, 2005.

[89] K. Johnsen, A. Raij, B. Lok, A. Stevens, and D. S. Lind. The validity of a virtual human experience for interpersonal skills education. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007.

[90] D. W. Johnson. Cooperativeness and social perspective taking. *Journal of Personality and Social Psychology*, 31:241–244, 1975.

[91] A. Kalra and K. Karahalios. TextTone: Expressing Emotion Through Text. *INTERACT 2005*, 3585:966, 2005.

[92] C. Keller and M. Duffy. I Said That? *TEACHING Exceptional Children*, page 37, 2005.

[93] B. Kerr. Thread Arcs: An Email Thread Visualization. In *Proceedings of the 2003 IEEE Symposium on Information Visualization*, page 27, 2003.

[94] J. Kim, D. Gunn, E. Schuh, B. Phillips, R. Pagulayan, and D. Wixon. Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM New York, NY, USA, 2008.

[95] A. Kitsantas and B. Zimmerman. Enhancing self-regulation of practice: the influence of graphing and self-evaluative standards. *Metacognition and Learning*, 1(3):201–212, 2006.

[96] B. Knerr and D. Lampton. An Assessment of the Virtual–Integrated MOUT Training System (V-IMTS). Technical Report ADA438315, United States Army Research Institute for the Behavioral and Social Sciences, 2005.

[97] B. W. Knerr, D. R. Lampton, G. A. Martin, D. A. Washburn, and D. Cope. Developing an after action review system for virtual dismounted infantry simulations. In *The InterserviceIndustry Training, Simulation & Education Conference (I/ITSEC)*, 2002.

[98] T. Koetter and M. Theus. Fortune: A system for interactive statistical graphics for time series. http://www.vr-web.de/~martin.theus/Fortune_JCGS.pdf. Self-published, December 2008.

[99] A. Kotranza and B. Lok. Virtual human + tangible interface = mixed reality human. an initial exploration with a virtual breast exam patient. In *IEEE Virtual Reality 2008*, pages 99–107, Reno, NV, March 2008.

[100] A. Kotranza, B. Lok, A. Deladisma, C. M. Pugh, and D. Lind. Mixed reality humans: Evaluating behavior, usability, and acceptability. *To Appear in IEEE Transactions on Visualization and Computer Graphics*, 2009.

[101] E. Krupat, R. Frankel, T. Stein, and J. Irish. The four habits coding scheme: Validation of an instrument to assess clinicians communication behavior. *Patient Education and Counseling*, 62(1):38–45, 2006.

[102] D. Lampton, J. Cohn, M. Endsley, J. Freeman, M. Gately, and G. Martin. Measuring Situation Aw reness for Dismounted Infantry Squads. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, 2005(-1), 2005.

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

[104] A. Lecuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet. Pseudo-haptic feedback: can isometric input devices simulate force feedback? *Proceedings of IEEE Virtual Reality 2000*, pages 83–90, 2000.

[105] J. Leigh, A. E. Johnson, T. A. DeFanti, M. Brown, and et al. A review of tele-immersive applications in the cave research network. In *IEEE Virtual Reality 1999*, pages 180–187, Houston, TX, March 1999.

[106] J. Lin, E. Keogh, and S. Lonardi. Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization*, 4(2):61–82, 2005.

[107] H. Liu, H. Lieberman, and T. Selker. Automatic Affective Feedback in an Email Browser. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2003.

[108] H. Liu, T. Selker, and H. Lieberman. Visualizing the affective structure of a text document. *ACM CHI'03 Extended Abstracts on Human Factors in Computing Systems*, April 2003.

[109] Lockheed Martin Information Systems. *Advanced Distributed Simulation Technology II (ADST II) Combat Synthetic Training Assessment Range (CSTAR) FAS Report*, 1997.

[110] A. MacEachren. *How Maps Work: Representation, Visualization, and Design.* Guilford Press New York, 1995.

[111] A. MacEachren, F. Boscoe, D. Haug, and L. Pickle. Geographic visualization: designing manipulable maps for exploringtemporally varying georeferenced statistics. In *Proceedings of IEEE Symposium on Information Visualization 1998*, pages 87–94, 1998.

[112] M. Macedonia. Games Soldiers Play. *IEEE Spectrum*, 39(3):32–37, 2002.

[113] J. Mackinlay, G. Robertson, and R. DeLine. Developing calendar visualizers for the information visualizer. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, pages 109–118. ACM New York, NY, USA, 1994.

[114] J. MacMillan, A. Alexander, S. Weil, B. Littleton, I. Aptima, M. Woburn, B. Roberts, W. Ferguson, J. Berliner, R. Tomlinson, et al. DARWARS: An Architecture That Supports Effective Experiential Training. Technical report, BBN Technologies, 2005.

[115] A. Manganas, M. Tsiknakis, E. Leisch, M. Ponder, T. Molet, B. Herbelin, N. Magnenat-Thalmann, D. Thalmann, M. Fato, and A. Schenone. The just vr tool: An innovative approach to training personnel for emergency situations using virtual reality techniques. *Journal on Information Technology in Healthcare*, 2(6):399–412, 2004.

[116] T. Manninen. Interaction Forms and Communicative Actions in Multiplayer Games. *Game Studies*, 3(1):2003, 2003.

[117] W. Mao and J. Gratch. The Social Credit Assignment Problem. *Proceedings of the Fourth International Working Conference on Intelligent Virtual Agents*, 2003.

[118] W. Mao and J. Gratch. Social Causality and Responsibility: Modeling and Evaluation. *Proceedings of the Fifth International Conference on Intelligent Virtual Agents*, 2005.

[119] P. May. Management Overview of the Intelligence and Electronic Warfare Tactical Proficiency Trainer. *2002 Spring Simulation Interoperability Workshop*, 2002.

[120] A. McCrindle and C. Christensen. The impact of learning journals on metacognitive and cognitive processes and learning performance. *Learning and Instruction*, 5(2):167–185, 1995.

[121] P. McDowell, R. Darken, J. Sullivan, and E. Johnson. Delta3D: A Complete Open Source Game and Simulation Engine for Building Military Training Systems. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, 2005, 2005.

[122] M. McGovern. The use of video in the self-evaluation of speech therapy students. *International Journal of Language & Communication Disorders*, 20(3):297–300, 1985.

[123] A. Mehrabian. *Nonverbal Communication*. Aldine-Atherton, Chicago, IL, 1972.

[124] A. Mehrabian. *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth, Belmont, California, 2nd edition, 1981.

[125] D. Miller and J. Thorpe. SIMNET: the advent of simulator networking. *Proceedings of the IEEE*, 83(8):1114–1123, 1995.

[126] M. Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970. Translated by Karl F. MacDorman and Takashi Minato.

[127] J. Morie, K. Iyer, D. Luigi, J. Williams, A. Dozois, and A. Rizzo. Development of a Data Management Tool for Investigating Multivariate Space and Free Will Experiences in Virtual Reality. *Applied Psychophysiology and Biofeedback*, 30(3):319–331, 2005.

[128] M. Morin, J. Jenvald, and M. Thorstensson. Computer-supported visualization of rescue operations. *Safety Science*, 35(1-3):3–27, 2000.

[129] T. Murphey. Why don't teachers learn what learners learn? Taking the guesswork out with Action Logging. In *English Teaching Forum*, volume 21, pages 6–10, 1993.

[130] C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

[131] NaturalPoint. Optitrack software. http://www.naturalpoint.com/optitrack/, November 2007. NaturalPoint, Inc.

[132] C. L. North. *A User Interface for Coordinating Visualizations Based on Relational Schemata: Snap-together Visualization*. PhD thesis, University of Maryland, May 2000.

[133] M. OReilly, M. OHalloran, J. Sigafoos, G. Lancioni, V. Green, C. Edrisinha, H. Cannella, and M. Olive. Evaluation of video feedback and self-management to decrease schoolyard aggression and increase pro-social behaviour in two students with behavioural disorders. *Educational Psychology*, 25(2):199–206, 2005.

[134] S. L. Oviatt, P. R. Cohen, M. Wang, and J. Gaston. A simulation-based research strategy for designing complex nl systems. In *ARPA Human Language Technology Workshop*, Princeton, NJ, 1993.

[135] D. F. Parkhurst. Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *BioScience*, 51(12):1051–1057, 2001.

[136] S. Paul, K. Dawson, J. Lanphear, and M. Cheema. Video recording feedback: a feasible and effective approach to teaching history-taking and physical examination skills in undergraduate paediatric medicine. *Medical Education*, 32(3):332–336, 1998.

[137] E. Payne and K. Dietrick. WARSIM Enters the Scene in Army Training. *CrossTalk The Journal of Defense Software Engineering*, 18(9):12–13, 2005.

[138] K. Perlin and D. Fox. Pad: an alternative approach to the computer interface. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 57–64. ACM Press New York, NY, USA, 1993.

[139] D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, 11(1):68–78, 2002.

[140] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. LifeLines: visualizing personal histories. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, 1996.

[141] A. Pusch, O. Martin, and S. Coquillart. Hemp hand-displacement-based pseudo-haptics: A study of a force field application. *Proceedings of IEEE 3D User Interfaces 2008*, pages 59–66, 2008.

[142] J. Quarles, S. Lampotang, I. Fischler, P. Fishwick, and B. Lok. A mixed reality system for enabling collocated after action review. In *7th IEEE and ACM International Symposium on Mixed and Augmented Reality - ISMAR 2008*, Cambridge, UK, September 2008.

[143] K. Räihä, A. Aula, P. Majaranta, H. Rantala, and K. Koivunen. Static visualization of temporal eye-tracking data. *Lecture Notes in Computer Science*, 3585:946–949, 2005.

[144] A. Raij, K. Johnsen, R. Dickerson, B. Lok, M. Cohen, M. Duerson, R. Pauley, A. Stevens, P. Wagner, and D. S. Lind. Comparing interpersonal interactions with a virtual human to those with a real human. *IEEE Transactions on Visualization and Computer Graphics*, 13(3), May 2007.

[145] A. Raij, K. Johnsen, R. Dickerson, B. Lok, M. Cohen, A. Stevens, T. Bernard, C. Oxendine, P. Wagner, and D. S. Lind. Interpersonal scenarios: Virtual ≈ real? In *IEEE Virtual Reality 2006*, pages 59–66, 2006.

[146] A. Raij, A. Kotranza, D. S. Lind, and B. Lok. Virtual experiences for social perspective-taking. In *To appear in Proceedings of IEEE Virtual Reality 2009*, Lafayette, Louisiana, March 2009.

[147] A. Raij and B. Lok. Ipsviz: An after-action review tool for human-virtual human experiences. In *IEEE Virtual Reality 2008*, Reno, NV, March 2008.

[148] R. Ramloll, C. Trepagnier, M. Sebrechts, and J. Beedasy. Gaze data visualization tools: opportunities and challenges. *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*, pages 173–180, July 2004.

[149] R. Rapee and K. Hayman. The effects of video feedback on the self-evaluation of performance in socially anxious subjects. *Behaviour Research and Therapy*, 34(4):315–322, 1996.

[150] R. Reeder, P. Pirolli, and S. Card. WebEyeMapper and WebLogger: tools for analyzing eye tracking data collected in web-use studies. In *Conference on Human Factors in Computing Systems*, pages 19–20, 2001.

[151] C. Reeve. Presence in virtual theater. *Presence: Teleoperators & Virtual Environments*, 9(2):209–213, 2000.

[152] E. Remolina, S. Ramachandran, R. Stottler, and W. Howse. Intelligent Simulation-Based Tutor for Flight Training. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, 2004(-1), 2004.

[153] S. Rezvan. The effects of metacognitive training on the academic achievement and happiness of Esfahan University conditional students. *Counselling Psychology Quarterly*, 19(4):415–428, 2006.

[154] D. RIDLEY. Self-Regulated Learning: the Interactive Influence of Metacognitive Awareness and Goal-Setting. *Journal of Experimental Education*, 60(4):293–306, 1992.

[155] A. A. Rizzo, D. Klimchuk, R. Mitura, T. Bowerly, J. G. Buckwalter, K. Kerns, K. Randall, R. Adams, P. Finn, I. Tarnanas, C. Sirbu, T. H. Ollendick, and S.-C. Yeh. A virtual reality scenario for all seasons: The virtual classroom. In *11th International Conference on Human-Computer Interaction*, Las Vegas, NV, July 2005.

[156] M. Rokeach. *The Open and Closed Mind: Investigations Into the Nature of Belief Systems and Personality Systems*. Basic Books, 1960.

[157] D. Roter, S. Larson, H. Shinitzky, R. Chernoff, J. Serwint, G. Adamo, and L. Wissow. Use of an innovative video feedback technique to enhance communication skills training. *Medical Education*, 38:145–157(13), February 2004.

[158] B. O. Rothbaum, L. Hodges, S. Smith, J. H. Lee, and L. Price. A controlled study of virtual reality exposure therapy for the fear of flying. *Journal of Consulting and Clinical Psychology*, 68(6):1020–26, 2000.

[159] K. Ryokai, C. Vaucelle, and J. Cassell. Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning*, 19(2):195–208, 2003.

[160] A. Santella and D. DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, pages 27–34, 2004.

[161] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, pages 336–343, College Park, Maryland 20742, U.S.A., 1996.

[162] B. Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[163] J. Shockley, R. Ford, M. Beebe, G. Lucha, and M. Johnson. The Joint Training Experimentation Program: Hotwash from the Second Demonstration. *04SSIW-132, Simulation Interoperability Workshop, April*, 2004.

[164] S. Silva and T. Catarci. Visualization of linear time-oriented data: a survey. In *Proceedings of the First International Conference on Web Information Systems Engineering*, volume 1, pages 310–319, 2000.

[165] E. Simon. *Masking Unmasked: Four Approaches to Basic Acting*. Palgrave Macmillan, 2003.

[166] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, 9(1):37–51, 2000.

[167] P. D. Sloane, R. Beck, V. Kowlowitz, A. M. Blotzer, L. Wang, L. Akins, F. White-Chu, and C. M. Mitchell. Behavioral coding for evaluation of medical student communication: Clarification or obfuscation? *Academic Medicine*, 79(2):162–70, February 2004.

[168] M. Smith and A. Fiore. Visualization components for persistent conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 136–143, 2001.

[169] R. Smith. IEWTPT System Design Review: Software Design. *US Army Simulation Training and Instrumentation Command*, 2001.

[170] R. Smith and G. Allen. After action review in military training simulations. In *WSC '94: Proceedings of the 26th Conference on Winter Simulation*, pages 845–849, San Diego, CA, USA, 1994.

[171] A. Steed, A. Brogni, and V. Vinayagamoorthy. Breaks in presence as usability criteria. In *Proc. of HCI International 2005, The 11th International Conference on Human Computer Interaction*, July 2005.

[172] W. Steptoe and A. Steed. High-fidelity avatar eye-representation. *IEEE Virtual Reality 2008*, pages 111–114, March 2008.

[173] V. Stevens. Making the most of a teachable moment: a smokeless-tobacco cessation intervention in the dental office. *American Journal of Public Health*, 85(2):231–235, 1995.

[174] S. Stoev and W. Straßer. A case study on interactive exploration and guidance aids for visualizing historical data. In *Proceedings of the IEEE Conference on Visualization*, pages 485–488. IEEE Computer Society Washington, DC, USA, 2001.

[175] C. Strapparava, A. Valitutti, and O. Stock. Dances with Words. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1719–1724, 2007.

[176] W. Swartout, J. Gratch, R. Hill, E. Hovy, S. Marsella, J. Rickel, D. Traum, and T. University Of Southern California Marina Del Rey Ca Inst For Creative. Toward virtual humans. *AI Magazine*, 27(2):96, 2006.

[177] T. Takada and H. Koike. Tudumi: Information Visualization System for Monitoring and Auditing Computer Logs. In *Proceedings on the IEEE Conference on Information Visualization*, pages 570–576, 2002.

[178] A. Tartaro and J. Cassell. Authorable virtual peers for autism spectrum disorders. In *Workshop on Language-Enabled Educational Technology at the 17th European Conference on Artificial Intelligence (ECAI06)*, 2006.

[179] K. Thiemann and H. Goldstein. Social stories, written text cues, and video feedback: Effects on social communication of children with autism. *Journal of Applied Behavior Analysis*, 34(4):425–446, 2001.

[180] K. R. Thórisson. Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people. In *1st ACM International Conference on Autonomous Agents*, pages 536–7, Marina del Rey, CA, 1997.

[181] K. R. Thórisson and J. Cassell. Why put an agent in a body: The importance of communicative feedback in human-humanoid dialogue. In *Lifelike Computer Characters '96*, Snowbird, UT, 1996.

[182] A. Unwin and G. Wills. Exploring Time Series Graphically. *Statistical Computing and Graphics Newsletter*, 2:13–15, 1999.

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

[184] A. Van Buren and S. Nowicki. Awareness of Interpersonal style and self-evaluation. *The Journal of social psychology*, 137(4):429–434, 1997.

[185] J. van Wijk and E. van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of the 1999 IEEE Symposium on Information Visualization*, page 4. IEEE Computer Society Washington, DC, USA, 1999.

[186] F. Viégas and J. Donath. Chat circles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: The CHI is the Limit*, pages 9–16. ACM New York, NY, USA, 1999.

[187] F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: portraying relationships from conversational histories. In *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 979–988, 2006.

[188] V. Vinayagamoorthy, A. Brogni, M. Gillies, M. Slater, and A. Steed. An investigation of presence response across variations in visual realism. *The 7th Annual International Presence Workshop*, pages 148–155, 2004.

[189] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater. Building expression into virtual characters. In *Eurographics 2006 State of the Art Reports*, Vienna, Austria, 2006.

[190] V. Vinayagamoorthy, A. Steed, and M. Slater. Building characters: Lessons drawn from virtual environments. In *Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, pages 119–26, Stresa, Italy, 2005.

[191] T. Watanabe, M. Ogikubo, and Y. Ishil. Visualization of respiration in the embodied virtual communication system and its evaluation. *International Journal of Human-Computer Interaction*, 17(1):p89 – 102, March 2004.

[192] M. Wattenberg. Sketching a graph to query a time-series database. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 381–382, New York, NY, USA, 2001. ACM.

[193] C. Weaver, D. Fyfe, A. Robinson, D. Holdsworth, D. Peuquet, and A. MacEachren. Visual Analysis of Historic Hotel Visitation Patterns. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[194] M. Weber, M. Alexa, and W. Müller. Visualizing Time-Series on Spirals. In *Proceedings of the IEEE Symposium on Information Visualization*, page 7, 2001.

[195] L. A. Wind, J. van Dalen, A. M. M. Muijtjens, and J.-J. Rethans. Assessing simulated patients in an educational setting: the masp (maastricht assessment of simulated patients). *Medical Education*, 38(1):39–44, 2004.

[196] D. Wooding. Fixation Maps: Quantifying eye-movement traces. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pages 31–36, 2002.

[197] D. Wuillemin, G. van Doorn, B. Richardson, and M. Symmons. Haptic and visual size judgements in virtual and real environments. In *The 1st Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pages 86–89, 2005.

[198] R. Xiong and J. Donath. Peoplegarden: creating data portraits for users. In *UIST '99: Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology*, pages 37–44, 1999.

[199] N. Yee and J. Bailenson. Walk a mile in digital shoes: The impact of embodied perspective-taking on the reduction of negative stereotyping in immersive virtual environments. *Proceedings of PRESENCE*, page 2426, 2006.

[200] N. Yee and J. Bailenson. The proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research*, 33(3):271–290, 2007.

[201] C. Zanbaka, P. Goolkasian, and L. F. Hodges. Can a virtual cat persuade you? the role of gender and realism in speaker persuasiveness. In *Conference on Human Factors in Computing Systems*, pages 1153–1162, Montreal, Quebec, Canada, 2006.

[202] C. Zanbaka, A. Ulinski, P. Goolkasian, and L. F. Hodges. Effects of virtual human presence on task performance. In *The International Conference on Artificial Reality and Telexistence*, pages 174 – 181, 2004.

[203] B. Zimmerman. Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, 25(1):3–17, 1990.

BIOGRAPHICAL SKETCH

Andrew Brian Raij was born in 1979 in Miami Beach, Florida. Andrew was raised in South Florida, where he graduated cum laude from the Ransom Everglades School in 1997. In 2001, he received a bachelor of science degree in computer science and a minor in English from Northwestern University. After Northwestern, he was a graduate student in computer science at the University of North Carolina at Chapel Hill (UNC-CH). At UNC-CH, he conducted research in the areas of computer graphics, computer vision, and projector-camera systems under the supervision of Henry Fuchs, Marc Pollefeys, and Herman Towles.

After completing a master's degree in computer science at UNC-CH in 2003, Andrew began a Ph.D. program at the University of Florida (UF) in computer engineering under the supervision of Dr. Benjamin Lok. He was awarded a UF Alumni Fellowship to support his research, which focuses on virtual human interfaces and their emerging use for interpersonal skills training. His dissertation explores leveraging interaction logs, visualization, and virtual environments to create enhanced after-action reviews of interactions between humans and virtual humans. His work has received significant recognition in both the fields of computer science and medicine with 14 articles published in leading journals and conferences, including a featured article in the May 2007 issue of *IEEE Transactions on Visualizations and Computer Graphics*. He is a member of the IEEE and ACM, and served as a panelist at IEEE Virtual Reality 2008. He is also a member of the Delta Epsilon Iota Academic Honor Society, Alpha Lambda Delta Academic Fraternity, and the National Society of Collegiate Scholars.

On September 21, 2008, Andrew married his college sweetheart, Emily. Andrew and Emily currently reside in Gainesville, Florida with their dog Ruby. Upon completion of his Ph.D. program, Andrew will continue his research as a postdoctoral researcher at the University of Florida.