Transformed Social Interaction in Collaborative Virtual Environments

Jeremy N. Bailenson

Department of Communication

Stanford University

Introduction

In this chapter, I first plan to briefly discuss definitions of immersive virtual reality. Next, I focus specifically on the idea of digital human representation in virtual reality, in particular the manners in which avatars are constructed and currently utilized during social interaction. I then present a theoretical paradigm called Transformed Social Interaction (TSI), a research paradigm that explores the ways that VR allows people to interact in ways not possible face to face, and review a number of published studies examining TSI as well as some new pilot data. I conclude by discussing what the future looks like in regards to the next step in VR in research and applications, and discussing potential ethical problems with TSI.

What is Immersive Virtual Reality?

A video game powered by a joystick is an extremely simple form of virtual reality. Both are digital environments in which a user can interact. However, there are a number of major distinctions between immersive virtual reality and a video game. We discuss two of them here.

The first distinction concerns *user tracking*. In a video game, a user's behavior and form are tracked in an extremely unnatural manner. In the game, behaviors such as walking, running, and arm movements occur as a result of an abstract action such as a button push or a joystick movement. The user's form is arbitrary and bears little resemblance to the actual face or body of the user. On the other hand, immersive virtual reality simulations employ naturalistic tracking. Sensitive equipment typically utilizing optical, magnetic, or mechanical sensors track a user's behaviors. In other words, if a user wants to walk through an immersive virtual world, instead of moving a joystick, she just walks through the physical world, and the system then renders (i.e., projects an image or sound of that behavior) that walking behavior in the physical world. Another way to express the idea of naturalistic tracking is that there is a one to one mapping of the user's movements in the physical world and the user's movements through the virtual world.

The second major distinction is the concept of *immersion* and immersive displays. Immersive virtual reality surrounds the user perceptually with sensory input (most commonly visual and auditory information). Instead of hearing sounds from a source localized in a single place, sounds emanate from various places in the digital environment. Likewise, visual input is rendered from anywhere in the user's visual field, as opposed to having to look at a computer monitor in a single place. Immersive virtual reality systems utilize equipment that renders sensory information stereoscopically (i.e., different information reaches each individual eye and ear) and spatially localized from anywhere in the users sensory field.

Many social scientists are using immersive virtual reality to study human behavior (see Blascovich et al., 2002 or Loomis, Blascovich, & Beall, 1999 for a review). In addition, a number of scholars are seeking to examine the concept of presence, the degree to which users feel they are in the virtual world as opposed to the physical world (see Lee, 2004, for a recent review of this work). This paper, however, will focus on the use of immersive virtual reality to join individuals from remote physical locations into the same virtual world.

Collaborative Virtual Environments

Collaborative Virtual Environments (CVEs) are simply virtual environments in which contain more than one user simultaneously. Users in CVEs interact with *avatars*, digital representations of one another (see Bailenson & Blascovich, 2002, for an explication of this concept). As Person A moves in Chicago, the tracking equipment measures all of his movements, gestures, expressions and sounds. Person B, in New York, already has a digital environment that contains the digital information necessary to render (i.e., create an on the fly, digital representation) three-dimensional model of the avatar of Person B as well as the information necessary to render whatever specific digital environment their avatars inhabit at that point in time. Person A's equipment then sends all of the tracking information over a network to Person B's rendering machine, which then renders all of those movements onto Person A's avatar. This bidirectional process—track the users' actions, send those actions over the network, and render those actions simultaneously for each user—occurs at an extremely high frequency (e.g., 60 times per second).

Compare a CVE to a traditional videoconference, depicted in Figure 1. In a videoconference, a camera records each user and then sends that stream of information over the network directly into the display of the other user. In other words, the brunt of the work is being performed by the network, which automatically sends every

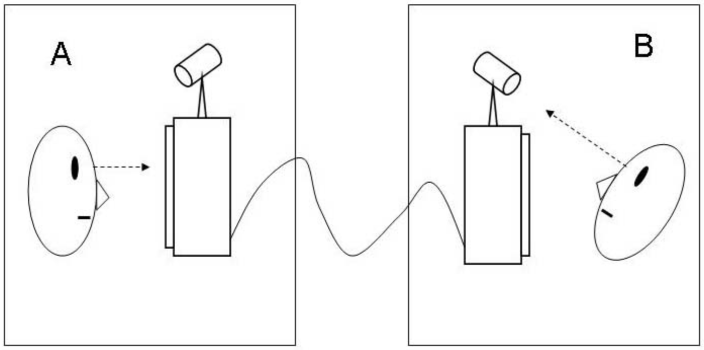

Figure 1: A schematic of a simple videoconference.

piece of information that each of the two cameras record.

Videoconferences tend to be relatively taxing on users for a variety of reasons. One of which is the classic inability to orient eye contact with traditional videoconferences. As Figure 1 demonstrates, a user can either look at the image of the other user (as A is doing) or she can look directly at the camera (as B is doing), but it is impossible to accomplish both of those tasks at once. The inevitable result is a feeling of disconnectedness due to the inability to match eye contact and other gestures. For example, in the above figure, Person A sees Person B looking right at his eyes. However, Person B cannot see the image of Person A, because she is looking at the camera, and even if she could, she would see an image of Person A looking at her nose, not her eyes. In this example, there are only two people and disconnectedness occurs. The problem only becomes exacerbated in videoconferences that involve members on three or more separate locations, for obvious reasons. There have been some notable attempts to correct this gesture disconnect problem in video conferences (see Yang & Zhang, 2002, for a review of these solutions), however these solutions typically involve some type of behavior tracking and digital rendering, similar in concept to a CVE.

In addition, even with high-bandwidth connections, there tends to be a delay between performed behavior and received behavior across the network. This is due to the system repeatedly sending every piece of information via an analog-like information stream. In other words, even if two people stand frozen like statues in a videoconference, the system still sends the same amount of information as if they were both talking, moving, and gesturing. Consequently, there are inherently delays in videoconferencing systems, and these delays can be extremely disruptive to the conversational flow and interactional synchrony (Kendon, 1977) in a conversation. Even as bandwidth increases, this problem is unlikely to be solved due to the corresponding increase in image resolution and sound quality.

On the other hand, consider the CVE depicted below in Figure 2. On the left panel of the diagram, three users in remote physical locations are being tracked and rendered. The right panel demonstrated the digital configuration in which the three users

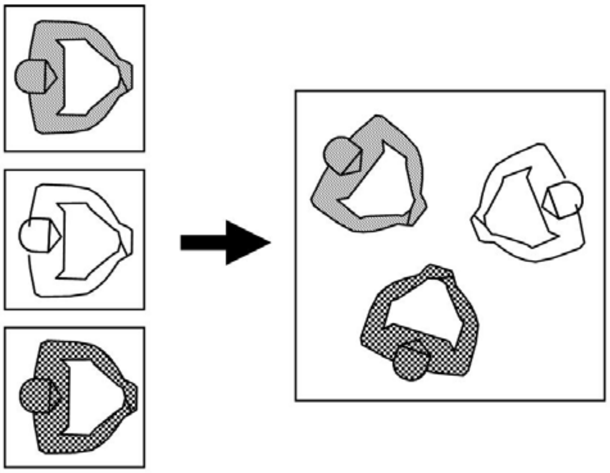

Figure 2: A schematic of a simple CVE.

see and hear one another. Each of the users has a digital image of the other two stored locally in their CVE system. The CVE receives digital tracking information about movements, gestures and other actions from the network; this information is a tiny fraction of the size of analog-like image and sound information. Consequently, there is virtually no delay at all in the transmission, because the only information that needs to be sent is a small stream of numbers that represents the current behaviors of each user. Furthermore, in a CVE, it is trivial to ensure that user's receive the proper eye-contact and connect on other gestures because the system can use algorithms to adjust incoming tracking information to keep the configurations set in the desired manner. Consequently the disconnect problem depicted in Figure 1 occurs extremely infrequently in CVEs.

Transformed Social Interaction Theory

Given that CVEs render the world separately for each user simultaneously, it is possible to break the normal physics of conversation and to render the interaction differently for each user at the same time. In other words, each CVE user sends the other users a particular stream of information that summarizes his or her current state of actions. However, that user theoretically can alter his stream of information in real time for strategic purposes. The theory of Transformed Social interaction (TSI; Bailenson, Beall, Loomis, Blascovich & Turk, 2004; Bailenson & Beall, 2004) examines the possibilities that these real-time transformations raise. TSI explores three dimensions for transformations during interaction.

The first dimension of TSI is transforming *sensory abilities*. These transformations compliment human perceptual abilities. One example of this transformation is to render 'invisible consultants', either algorithms or human avatars who are only visible to particular members of the CVEs. These consultants can either provide real-time summary information about the attentions and movements of other interactants (information which is automatically collected by the CVE) or can scrutinize the actions of the user herself. For examples, teachers using distance learning applications can utilize automatic registers that ensure they are spreading their attention equally towards each student.

The second dimension is *situational context,* transforming the spatial or temporal structure of a conversation. For example, each user in the CVE can optimally configure the geographical setup of the room. Using the distance learning paradigm, every single student in a class of twenty can sit directly in front of the virtual blackboard, and perceive the rest of the students as sitting behind him. Furthermore, by altering the flow of rendered time in a CVE, users can implement strategic uses of “pause” and “rewind” during a conversation in attempt to increase comprehension and productivity.

The third dimension of TSI is *self representation,* the strategic decoupling the rendered appearance or behaviors of avatars from the actual appearance or behavior of the human driving the avatar. Because interactants can modulate the flow if information dictating the way their avatars are rendered to others, that rendering can deviate from the actual state of the user. In the distance learning paradigm, it could be the case that some students learn better with teachers who smile while some learn better with teachers with serious faces. In a CVE, the teacher can render himself differently to those students, tailoring his facial expressions to each student in order to maximize their attention and learning.

In sum, using TSI, users can strategically alter aspects of the conversation. Previous work has discussed the ability of people to use limited forms of TSI during face-to-face interaction (Bailenson & Beall, 2004), such as applying makeup, plastic surgery, and automatic nonverbal mimicry such as "the chameleon effect" (Chartrand & Bargh, 1999). However there is no doubt that the advent of CVEs allows for a plethora of TSI strategies on a much larger scale. In the following section, we review empirical work that has examined TSI in CVEs.

Empirical Investigation of TSI

Multilateral Perspective Taking

A CVE is rendered from scratch 60 times per second for each of the interactants. Normally, this is done by determining where each interactant is standing and looking, and rendering the scene appropriately given that information. However, an interactant in theory can render her sensory point of view from any single place in the room. In other words, it is possible for Person B to disconnect her area of perception from the area in which Person A perceives her. Figure 3 illustrates this transformation.

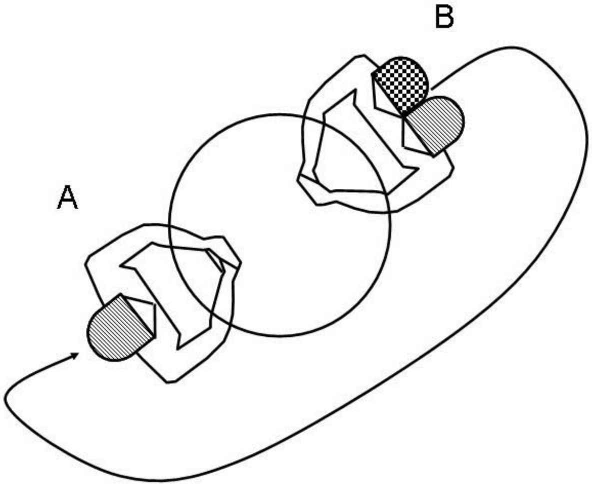

Figure 3: Person B takes on multilateral perspectives: she can experience the CVE her own perspective and the perspective of Person A at the same time.

In the above Figure, Person B is implementing a multilateral perspective. Specifically, she is choosing to adopt the sensory perspective of Person A during the conversation. In other words, she has left her own point of view and become a passenger to Person A, by rendering not a digital world that is contingent on her own movements, but instead a digital world that is contingent on Person A's movements. As a result, she sees herself in real-time from behind the eyes of her conversational partner. Either by shifting her entire field of view to the spatial location of other avatars in the interaction, or by popping up "field of view windows" in corners of the virtual display, an interactant can unobtrusively occupy the home space of any avatar in the CVE.

Current research (Gehlbach, Bailenson, Yee, & Beall, 2004) is examining multilateral perspectives in a negotiation scenario inside a CVE. Previous work has used either role playing (Davis, Conklin, Smith, & Luce, 1996) or observational seating arrangements (Taylor & Fiske, 1975) to cause subjects to take on the perspectives of others in a conversation, demonstrating more efficient and effective interactions. Equipping an interactant with the real-time ability to see one's avatar from another point of view should enhance these previous findings. In our current work in progress, we are predicting to find more cooperative solutions in simulations in which negotiators can occupy the field of view of their opponents.

We also are currently utilizing this perspective-shifting tool as a way to better implement diversity training. The general idea is to create some type of simulation that depicts a diversity-related interpersonal event, and to then use the simulation to put subjects in the perspective of all parties involved in the diversity event. Recent work by

Yee and Bailenson (2004) utilizes the “virtual mirror”. In these studies, subjects arrive into the lab, enter into an immersive virtual reality simulation, and walk around a simulated room. The main feature in the simulated room is a virtual mirror. The mirror shows the reflection of their avatar; the avatar moves and gestures in the mirror as the experimental subject moves and gestures in the physical room. However, this is not a normal mirror, in the sense that we can render their mirror image in whatever form we choose—skin color, gender, height and attractiveness are all flexible parameters in the virtual mirror.

Our research paradigm entails having subjects stand and gesture in front of the virtual mirror for about ninety seconds, watching their reflection, which is an image of their opposite gender or race. Next, they turn around from the mirror and interact with two other avatars that have entered the CVE. We then measure their behavior towards the other avatars: the amount of personal space they leave between themselves and the others in the CVE (Bailenson, Blascovich, Beall, & Loomis, 2003), the amount of direct eye contact they make with others in the CVE, the amount of time they spend talking in the interaction, and a number of subjective measures concerning the subjects’ overall confidence and task performance when they are seen by others as wearing a mismatched avatar. Allport (1954) is most credited with the “contact hypothesis”, the notion that putting people from different social groups together reduces prejudice. Using the virtual mirror, we are exploring the “hyper-contact hypothesis”, namely that being forced to wear the face and body of a member of the opposite social group should produce an even further reduction of prejudice.

Non-zero Sum Gaze

Another TSI tool is *non-zero-sum gaze (NSZG)*: directing mutual gaze at more than a single interactant in a CVE at once. Previous research has demonstrated that eye gaze is an extremely powerful tool for communicators seeking to garner attention, be persuasive and instruct (see Segrin, 1993, for a review on this topic). People who use mutual gaze increase their ability to engage an audience as well as to accomplish a number of conversational goals.

In face-to-face interaction, gaze is zero-sum. In other words, if Person A looks directly at Person B for 65 percent of the time, it is not possible for Person A to look directly at Person C for more than 35 percent of the time. However, interaction among avatars in CVEs is not bound by this constraint. The virtual environment as well as the other avatars in the CVE is individually rendered for each interactant locally. As a result, Person A can have his avatar rendered differently for each other interactant, and appear to maintain mutual gaze with both B and C for a majority of the conversation, as Figure 4 demonstrates.

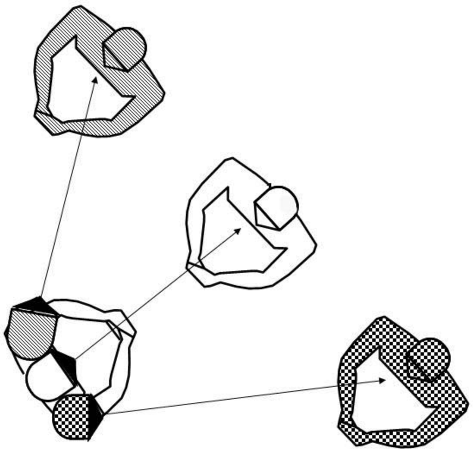

Figure 4: A schematic illustration of non-zero-sum gaze. Each interactant on the right perceives the speaker on the left gazing directly at him or her.

NZSG allows interactants to perpetuate the illusion that they are looking directly at each person in an entire roomful of interactants. Three separate projects (Bailenson, Beall, Blascovich, Loomis, & Turk, 2004; Beall, Bailenson, Loomis, Blascovich & Rex, 2003, Guadagno et al. 2004] have utilized a paradigm in which a single presenter read a passage to two listeners inside a CVE. All three interactants were of the same gender, wore stereoscopic, head-mounted displays, and had their head movements and mouth movements tracked and rendered, and the presenter's avatar either looked directly at each of the other two speakers simultaneously for 100 percent of the time (augmented gaze) or utilized normal, zero-sum gaze. Results across those three studies have demonstrated three important findings: 1) participants never detected that the augmented gaze was not in fact backed by real gaze, 2) participants returned gaze to the presenter more often in the augmented condition than in the normal condition, and 3) participants (females to a greater extant than males) were more persuaded by a presenter implementing augmented gaze than a presenter implementing normal gaze.

NZSG should be a powerful tool in future computer-mediated communication. Walther (1996) describes how interactions conducted via digital representations can be *hyperpersonal*: more intimate and intense than face-to-face interaction. The current findings extend the notion of hyperpersonal interaction to the nonverbal realm. For applications such as distance education, sales, online chatting and dating, utilizing computer guided gaze should have a high impact on social interaction.

Digital Chameleons

Previous research on face-to-face interaction has shown that people are highly influenced by others who mimic their language (van Baaren, Holland, Steenaert, & van

Knippenberg, 2003) or their gestures (Chartrand & Bargh, 1999) during social interaction. In this latter study, an interviewee who mimicked the posture and sitting position of the interviewer during the interaction was seen more favorably than one who did not perform the mimicry. Moreover, recent research has extended these findings to computer agents: voice synthesizers that mimic vocal patterns (Suzuki, Takeuchi, Ishii, & Okada, 2003) as well as embodied agents in immersive virtual reality that mimic nonverbal behavior (Bailenson & Yee, 2005).

In the study by Bailenson & Yee (2005) a realistic looking *embodied agent* (i.e., virtual human in a CVE that is controlled by a computer algorithm, not by another human) administered a three-minute verbal persuasive message. For half of the subjects, the agent was a digital chameleon, in that its head movements were an exact mimic of the subject’s head movements at a four second delay. For the other half, the agent’s head movements was a playback of another subject’s head movements. Results demonstrated three important findings: 1) participants rarely detected their own gestures when those gestures were utilized by the agents, 2) participants were more persuaded by the agent and liked the agent more in the mimic condition than in the recorded condition, and 3) participants actually looked at the agent more in the mimic condition than in the recorded condition.

In CVEs, as well as in many other forms of computer-mediated communication, people interact with one another through digital representations. Consequently, the potential for implementing digital chameleons in a variety of media remains a distinct possibility, one that is in many ways more powerful than face-to-face mimicry. For example, in computer-mediated communication, algorithms can be indiscriminately and constantly applied to mimic peoples' behaviors. Once mimics are utilized on an algorithmic level, the potential to utilize more fine grained strategies such as probabilistic mimics (e.g., Person A's avatar mimic's Person B's actions for only part of the time), rule-based mimics (e.g., Person A's avatar mimic's Person B's actions only when Person B is talking) and scaled mimics (e.g., Person A's avatar's head movements only move half as much as Person B's movements during the mimic) become almost trivial to employ. Such a systematic application of various levels of mimicry is difficult to implement in face-to-face conversation. Consequently, CVEs provide unique potential for social scientists to study contingency and mimicry during interaction, for example systematically exploring notions of interactional synchrony (Kendon, 1977).

Conclusions and Implications

In sum, the prevalence of research on TSI is growing steadily over the past five years. Across studies, two general patterns of results have been consistently emerging. First, TSI strategies are detected surprisingly infrequently during social interaction. Second, avatars that utilize TSI gain unique social influence in the interaction compared to avatars that do not utilize the transformation strategies.

In future work, we plan on exploring TSI phenomenon more thoroughly, including TSI in non-immersive, non-virtual settings. We have recently completed a series of studies examining *identity capture* (Bailenson, Garland, Iyengar, & Yee, 2004). We presented undergraduates with an image of an unfamiliar political candidate who was ostensibly running for office in California. The two-dimensional photograph of the candidate was either morphed with a photograph of the undergraduate filling out the questionnaire about the candidate or not. Results demonstrated that, under a variety of conditions, candidates that captured aspects of the specific subjects' facial structure gained advantage and was more likely to receive the subjects' votes. This result is not surprising, given that a variety of research indicates that people show affinity towards things that resemble themselves (see Baumeister, 1998, for a review). However, the truly surprising finding was that less than five percent of the subjects detected their own face in the photograph, even when the morphs captured up to 40 percent of the subjects' original photographs.

The fact that TSI can be effective in non-immersive settings is somewhat alarming. When people enter an immersive virtual reality simulation, the expectation or at least the accepting of some degree of foul play in terms of veridical rendering of other people's behavior is most likely inevitable. However, when viewing two-dimensional video feeds, images on web sites, voices enhanced by digital algorithms on cell phones, other players in online video games (Yee, 2004) and text in chat rooms, we may not be so rigorous in our skepticism concerning the authenticity of form and behavior. The potential for using TSI for abuse in all forms of digital communication certainly warrants attention.

The Orwellian themes behind this research paradigm are quite transparent. TSI tools such as identity capture, augmented gaze, and digital mimicry certainly would be better left out of the hands of advertisers, politicians, and anyone else whose may seek to influence people. On a more basic level, not being able to trust the very pillars of social interaction—what a person looks like and how they behave—presents interactants in a difficult position. On the other hand, is TSI fundamentally different from plastic surgery, makeup, self-help books and white lies?

Certainly the potential ethical concerns of TSI largely vanish if one assumes that all interactants in a CVE are aware of the potential for everyone to rampantly use these transformations. However, then the possibility becomes that TSI will cause CVE technology as a whole to become completely useless; why bother to use a communication device in which it is not possible to trust any of the actions of the other interactants? As computer-mediated communication becomes more advanced and prevalent, it will be fascinating to monitor the progress of TSI strategies as well as technology designed to detect and foil the non-veridical rendering of appearance and behaviors. In the meantime, TSI and CVEs present spectacular opportunities for social scientists studying nonverbal behavior, computer-mediated communication, and digital human representation.

References

Allport, G. (1954) *The Nature of Prejudice*, Reading, MA: Addison-Wesley.

Bailenson, J.N. & Beall, A.C. (2004, in press). Transformed Social Interaction: Exploring the Digital Plasticity of Avatars. In Schroeder, R. & Axelsson, A.'s (Eds.), Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments, Springer-Verlag.

Bailenson, J.N., Beall, A.C., Blascovich, J., Loomis, J., & Turk, M. (2004). Non-Zero-Sum Gaze and Persuasion. *Paper presented in the Top Papers in Communication and Technology session at the 54th Annual Conference of the International Communication Association,* New Orleans, LA.

Bailenson, J.N., & Blascovich, J. (2004, in press) Avatars. *Encyclopedia of Human-Computer Interaction*, Berkshire Publishing Group, 64-68.

Bailenson, J.N., Blascovich, J., Beall, A.C., & Loomis, J.M., (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29. 1-15.

Bailenson, J.N., Beall, A.C., Loomis, J., Blascovich, J., & Turk, M. (2004). Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments. *PRESENCE: Teleoperators and Virtual Environments*, 13(4), 428-441.

Bailenson, J. N., Garland, P., Iyengar, S., & Yee, N (2004). The effects of morphing similarity onto the faces of political candidates. Political Psychology. Manuscript under review.

Bailenson, J., Yee, N. (2005, in press). Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments, *Psychological Science*.

Baumeister, R.F. (1998). The self. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), Handbook of social psychology (4th ed.; pp. 680-740). *New York: McGraw-Hill.*

Beall, A. C., Bailenson, J. N., Loomis, J., Blascovich, J., & Rex, C. (2003). Non-zero-sum mutual gaze in immersive virtual environments. *Proceedings of HCI International 2003, Crete.*

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13,* 146-149.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: A merging of self and other. *Journal of Personality and Social Psychology, 70,* 713-726.

Gehlbach, H., Bailenson, J.N., Yee, N., & Beall, A.C. (2004). Perspective taking and negotiation in collaborative virtual environments. (draft in progress).

Guadagno, R. E., Bailenson, J. N., Beall, A. C., Dimov, A., & Blascovich, J. (2004). Non-zero sum gaze and the cyranoid: The impact of non-verbal behavior on persuasion in an immersive virtual environment. (draft in progress).

Kendon, A. (1977). *Studies in the Behavior of Social Interaction*. IU: Bloomington, Indiana.

Lee, K., M. (2004). Presence, Explicated. *Communication Theory, 14,* 27-50.

Loomis, J.M., Blascovich, J.J., & Beall, A.C. (1999). Immersive virtual environments as a basic research tool in psychology. *Behavior Research Methods, Instruments, and Computers, 31(4),* 557-564.

Segrin, C. (1993). The effects of nonverbal behavior on outcomes of compliance gaining attempts. *Communication Studies, 44*, 169-187.

Suzuki, N., Takeuchi, Y., Ishii, K., & Okada, M. (2003). Effects of Echoic Mimicry Using Hummed Sounds on Human-Computer Interaction, *Speech Communication*, *40*, 559–573.

Taylor, S.E. & Fiske, S.T. (1975) Point of view and perception so causality, *Journal of Personality and Social Psychology*, *32,* 439-445.

Yang, R., & Zhang, Z. (2002). Eye Gaze Correction with Stereovision for Video-Teleconferencing, *Proceedings of the Seventh European Conference on Computer Vision* (ECCV 2002) May 27 – June 2, 2002, Copenhagen, Denmark.

Yee, N. (2004, in press). The Psychology of MMORPGs: Emotional Investment, Motivations, Relationship Formation, and Problematic Usage. In R. Schroeder & A. Axelsson (Eds.), *Social Life of Avatars II*. London: Springer-Verlag.