A Longitudinal Study of Task Performance, Head Movements, Subjective Report, Simulator Sickness, and Transformed Social Interaction in Collaborative Virtual Environments

Jeremy N. Bailenson

Nick Yee

Department of Communication

Stanford University

In Press, *PRESENCE: Teleoperators and Virtual Environments*

Jeremy Bailenson

Department of Communication, Stanford University

Stanford, CA, 94305-2050

Bailenson@stanford.edu

Abstract

Empirical research on human behavior in collaborative virtual environments (CVEs) is in its infancy. Historically, one of the more valuable tools social scientists have used to evaluate new forms of media is longitudinal studies that examine user behavior over an extended period of time. In the current study, three triads of participants came to the lab for 15 sessions over a ten week period to collaborate for approximately 45 minutes per session. We examined nonverbal behavior, task performance on verbal tasks, and subjective ratings of presence, copresence, simulator sickness, and entitativity over time. Furthermore, we examined two types of transformed social interaction: nonverbal mimicry and facial similarity. Results demonstrated substantial differences in task performance, subjective ratings, nonverbal behavior, and simulator sickness over time as participants became familiar with the system. Furthermore, transforming avatar appearance to increase facial similarity sometimes improved task performance. We discuss implications for research on CVEs.

Introduction

Researchers in communication, computer science, and psychology are beginning to dedicate serious research efforts towards developing and understanding collaborative virtual environments (CVEs). CVEs are systems which allow geographically separated individuals to interact with one another via digital avatars which are rendered to look and behave consistently with their human counterparts, whose behaviors are measured by various forms of tracking equipment in real-time. Many researchers are optimistic that CVEs will have the unique ability to overcome problems associated with traditional teleconferencing equipment (Lanier, 2001) such as transmission delays and eye-contact problems resulting from the difficulty of dynamic camera placement.

Given that the study of CVEs is a relatively new endeavor, existing research on user behavior in CVEs is in its infancy. While a number of empirical studies examine task performance and quality of experience in CVEs, without exception, all of these studies examine experimental participants for whom the experiment itself is largely the first time they have experienced a CVE. Even the few studies which have utilized people with CVE experience did not systematically track those users' learning and adjustments to the extremely novel technology over time.

CVE studies typically focus on how presence, co-presence and task performance are influenced by particular aspects of the CVE, such as embodiment (Benford, Bowers, Fahlen, Greenhalgh, & Snowdon, 1995), awareness (Benford et al., 1995; Dourish & Bellotti, 1992), avatar realism (Casanueva & Blake, 2000), the presence of eye contact (Bailenson, Beall, & Blascovich 2002; Garau et al., 2003) or whether the CVE is multiscale or not (Zhang & Furnas, 2002). Others have focused on social processes in CVEs, for example, formation of trust in persistent online virtual worlds (Schroeder & Axelsson, 2000; Yee, in press) and how immersion might confer leadership (Steed, Slater, Sadagic, Bullock, & Tromp, 1999). Also well-known is the COVEN project that explored more technical requirements and scalability of CVEs (Normand et al., 1999).

Longitudinal research in the laboratory, that is, empirical studies which track users over extended periods of time, is rare in behavioral research largely due to the laborious nature of bringing the same people back to the lab over and over again. For example, until the seminal work of Termin (1916), people with high intelligence were regarded by the general populace as social outcasts who could not form healthy relationships with others. Only after Termin followed the social development of gifted children over time did he prove with longitudinal data that not all smart people are destined to spend their lives with only books as acquaintances. There are countless instances of longitudinal data being essential to answering behavioral science questions, especially in regards to evaluation of new media use (Huesmann, Lagerspetz, & Eron, 1984).

Consequently, in the current work, we embarked upon a longitudinal study of user behavior in CVEs. We believed that longitudinal research examining CVEs would be especially important given the novelty of the equipment and the difficulty in adjusting to immersion in a single session, and predominantly were interested in two important issues. First, we sought to examine the way in which human behavior changes over time with CVE use and experience (e.g., IJsselsteijn et al., 1997). Second, we manipulated various types of transformed social interaction (TSI) and examined the effect of TSI on human behavior over time.

TSI is a research paradigm (Bailenson & Beall, in press; Bailenson, Beall, Loomis, Blascovich, & Turk, 2004) that examines the disjoint between human characteristics and behaviors that exist in physical space and the characteristics and behaviors that are rendered to others in a CVE. Because behaviors are tracked and rendered in CVEs, as opposed to directly transmitted via an analogue-type of information stream, interactants have the ability to filter, augment, or block their own behaviors from the eyes and ears of their conversational partners. Previous research on TSI has examined the effects of being able to look directly in the eye of more than one other person at once (Beall, Bailenson, Loomis, Blascovich, & Rex, 2003), mimicking the head movements of other interactants (Bailenson, Beall, Blascovich, Loomis, & Turk, 2004; Bailenson & Yee, 2005), and morphing the faces of interactants to absorb facial features of conversational opponents (Bailenson, Garland, Iyengar, & Yee, 2005). However, oftentimes TSI is detrimental in terms of conversational flow and interactional synchrony (Kendon, 1977), because natural movements are being replaced with algorithmic ones, destroying sensitive interplay between verbal and nonverbal cues. Consequently, we wanted to trace the adjustment to TSI algorithms over time.

In the current study, we sought to further explore the notion of using TSI specifically to instill similarity among team members. There is a wealth of research that indicates that teamwork functions better when team members are “in synch” with one another. In other words, when team members are more similar to one another, work proceeds more effectively. This effect occurs when similarity is defined in terms of demographics (Kirkman, Tesluk, & Rosen, 2004) or in terms of personality (Reeves & Nass, 1996). There is ample reason to believe that TSI can be used to accelerate this process of building team familiarity, comfort and similarity. We manipulated two types of TSI: behavioral team similarity and visual team similarity.

To accomplish behavioral similarity, we implemented a head-movement mimic algorithm. In other words, for each interactant in the CVE, the other two interactants' avatars in the CVE mimicked his or her head movements (regardless of the actual head movement behavior of those other two interactants). In this mimic condition, all three participants saw their team members mimic them simultaneously with a four second delay. Previous research has shown that people are more influenced by other people who mimic their language (van Baaren, Holland, Steenaert, & van Knippenberg, 2003) or their gestures (Chartrand & Bargh, 1999) than those that do not mimic them during social interaction. Moreover, this trend also occurs with digital 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).

We also examined visual similarity. All interactants within the CVE had their avatar constructed to be a photographically realistic analog of their own head and face. Research in social psychology has demonstrated large-scale effects of similarity on social influence. An individual judged more similar to a given person (compared to a less similar individual) is considered more attractive (Berscheid & Walster, 1974; Shanteau & Nagy, 1979), persuasive (Byrne, 1971), is more likely to receive political support (Bailenson, Garland, Iyengar, & Yee, 2005) and is more likely to elicit altruistic helping behavior in a dire situation (Dovidio, Gaertner, Anastasio, & Sanitioso, 1992). Consequently, in the visual similarity condition, for each interactant in the CVE, the digitized face of an individual interactant was used as the virtual face for both remaining interactants in the group of three. In other words, interactant A saw the avatars of interactant B and C as "wearing" the identical face of interactant A.

It is important to note that the notion of team similarity is very complex. Well-functioning teams have a wide range of both convergent and divergent attributes among team members. In the current work, we only examine the possibility of manipulating superficial, surface similarity among team members as a way to improve performance.

Method

Participants

Nine undergraduate students (six female, three male) from an introductory communication course participated in the study for partial course credit. None had used immersive virtual reality more than once before. None of the participants were pregnant or had epilepsy, and all were native speakers of English. The three male subjects formed one group; the six female subjects were randomly assigned into two additional groups of three.

Design

We manipulated one independent variable: TSI condition, which had three levels: *normal*, *face similarity*, and *mimic*. In the normal condition, participants saw one another wearing the appropriate (i.e., their own) faces and gesturing veridically. In the face similarity condition, each participant saw the other two participants wearing his or her face but gesturing veridically. In other words, from the point of view of participant C, participants A and B gestured normally but each wore the visual face model of C. In the mimic condition, each participant saw the other two wearing their correct face models (see Figure 3), but each of them mimicked his or her head movements at a four second lag. In other words, participants A and B gestured with the head movements of C, but each wore the appropriate face. It is important to note that the field of view of the head-mounted display prevented any participant from seeing both of the other participants mimicking him simultaneously.

Each group collaborated with one another for fifteen trials. The trials were spread across a ten week academic quarter. Appendix A demonstrates the timetable for each group, and demonstrates that each group participated in each TSI condition five times. For each of the fifteen trials[1], each group received one of the three TSI conditions; in essence the TSI manipulation was a within-trial design. This term is used (as opposed to within-subject) because we use trial as a random factor in our statistical analysis instead of subject.

Procedure

Before the experiment began, all nine participants arrived at the laboratory and met one another face-to-face. They then had the procedure thoroughly explained, including specifications about the virtual reality equipment as well as the instructions on how to carry out the specific problem-solving tasks. Participants were instructed that they would be scored equally for quantity and accuracy of their responses. Furthermore, they had photographs taken of their faces to allow the construction of realistic avatars for use over the next ten weeks. During this pre-experiment meeting we explicitly described the three different experimental conditions and showed them the CVE. Upon leaving this meeting they were all aware that they would be mimicked in certain trials, that other people would wear their own face in certain trials, and that they could only see one other participant in any given field of view.

For each experimental session, one group of three participants arrived at the laboratory, filled out a consent form, and then each participant sat at an immersive virtual reality station. The stations were in opposite corners of a large room, and participants could not easily see one another if they were to take off their HMDs. First they engaged in a series of problem solving tasks while immersed in virtual reality. Then, they went to a separate room and filled out a number of self report questionnaires.

Problem solving tasks. Once the participants had successfully donned their equipment, each entered a CVE in which the other two participants were sitting around a table. They then began to perform three types of problem solving tasks in an order specified according to the counterbalancing scheme depicted in Appendix A. The tasks were administered verbally by the *operator*, an experimenter who could be heard via the audio system but had no visual presence inside the CVE. In between each set of tasks, participants took a mandatory three minute break in which they removed their HMD and audio equipment. A full list of the stimuli appears in Appendix B.

The first task was to perform three iterations of twenty questions, a task used previously to assess performance in CVEs (Bailenson, Beall, & Blascovich, 2002). During the twenty questions task, the operator would moderate the trials for each designated word. Members of the group would then deliberate about a “yes/no” question to ask that would narrow down the scope of what the word might be. When they agreed upon an appropriate question, one of the members would provide an “official” question, and the operator would answer it. For each iteration, the group had 20 questions or five minutes to determine the word. Providing corrective feedback for the participant is inherent in the 20 questions task.

The second task was to perform three iterations of the reverse remote association task, also used previously to assess performance in CVEs (Hoyt & Blascovich, 2003). For each iteration, participants received a word from the operator and were instructed to generate three words that were related to that one word. For example, given the word "ball", participants could generate the word "party", "bounce", and "bearing". Participants were told to keep two goals in mind equally when generating triads: 1) come up with as many as possible, and 2) come up with ones that were as creative as possible. We gave them some examples of uncreative ones (e.g., "basket", "base", and "foot") as well as more creative ones, in which individual members of the triad related to the target in different semantic manners. For each iteration, the group had five minutes to generate as many triads as possible, and during each experimental trial, a group received three separate RAT games with different words. There was no corrective feedback given by the operator about their responses.

The third task was to solve a single insight problem. Participants read the problem via three dimensional text projected inside the virtual world and could reread the problem by asking the operator. The problems were primarily taken from Weisberg (1995), and were designed foster collaborative thought in order to produce a solution. In other words, unlike a "hill-climbing" problem, in which one nears the solution in proportion to the amount of thought dedicated to the problem (i.e., filling in a crossword puzzle), solving insight problems require creativity, and regardless of the amount of work performed, one gets no closer to the solution without a flash of insight. Participants worked until they solved the problem or a ten minute time limit was up. There was no corrective feedback given by the operator about their responses.

Across these three types of tasks, participants spent about 35-40 minutes total immersed during each experimental trial. After completing the third task they removed the HMD, went to a separate room, and filled out a number of self-report questionnaires.

Self report tasks. While in a separate room from the immersive virtual reality equipment, participants sat down at a desk and completed a number of questionnaires. These measures are depicted in full in Appendix C.

The first was a four-item *Presence* questionnaire, designed to measure how immersed participants were in the virtual world, as opposed to the physical world. Cronbach's alpha, a measure commonly used to assess reliability of scales, was .75.

The second was a five-item, *Copresence* questionnaire, designed to measure how human-like and socially relevant the other avatars in the room were. This questionnaire was used previously in CVE research (Bailenson et al., 2002; Bailenson, Blascovich, Beall, & Loomis, 2003), and in this study Cronbach's alpha was .85.

The third was a ten-item, *Entitativity* questionnaire that measured how cohesive the group of three individuals were. In other words, a tightly-knit, well-functioning group would be high in entitativity, while a non-cohesive collection of disjointed individuals would be low. Items from this questionnaire were based on previous work (Maner et al., 2002), and in this study Cronbach's alpha was .85.

The fourth was a sixteen item simulator sickness questionnaire (Kennedy, Lane, Berbaum, & Lilienthal, 1993). Participants marked the extent to which they were experiencing various aftereffects on a scale from one to four with higher numbers indicating greater severity. Cronbach's alpha was .71 for this measure.

Finally, after each session, participants wrote an open-ended paragraph describing their experience that day in virtual reality. On average, the total self report session took about fifteen minutes.

Materials and Apparatus. Perspectively correct stereoscopic images were rendered by a 1700 MHz Pentium IV computer with an NVIDIA GeForce FX 5200 or 5950 graphics card, and were updated at an average frame rate of 60 Hz. The simulated viewpoint was continually updated as a function of the participants' head movements, which were tracked by a three-axis orientation sensing system (Intersense IS250, update rate of 150 Hz). The system latency, or delay between a participant's head movement and the resulting concomitant update in the HMD's visual display was 45 ms maximum. The software used to assimilate the rendering and tracking was Vizard 2.14.

Figure 1 shows a participant donning the equipment. Participants wore a Virtual Research V8 stereoscopic head mounted display (HMD) that featured dual 680 horizontal by 480 vertical pixel resolution panels that refreshed at 60 Hz, or an nVisor SX HMD that featured dual 1280 horizontal by 1024 vertical pixel resolution panels that refreshed at 60 Hz. On both types of HMD, the display optics presented a visual field subtending approximately 50 degrees horizontally by 38 degrees vertically.

Figure 1: A participant wearing the HMD.

Participants wore microphones and we used custom, real-time audio sampling software to measure the instantaneous speech sound levels captured near each participant's mouth. When the amplitude was over a certain threshold, their avatars opened their mouths to indicate speech. We sampled the microphone amplitude at a rate of 20 Hz. Furthermore, each participant had speakers placed near his or her station in order to clearly hear the utterances of the other two group members. The speech was non-spatialized, in that the sound did not emanate from the exact digital space containing the speaking avatar.

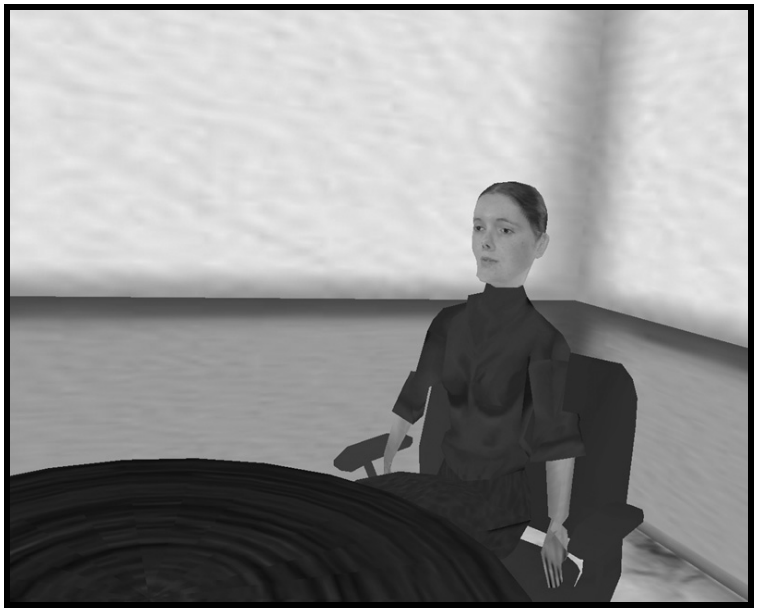


Figure 2: A participants' view of another participant's avatar.

Figure 2 demonstrates a sample point of view for a given participant. It was not possible for a participant to see both of the other two participants simultaneously; this setup was chosen to encourage head movements, which were easier to track and record than eye movements given our apparatus. The avatar faces were constructed from a series of photographs with software by 3dMeNow; previous research has indicated that these digital models are extremely high in similarity (both objectively and psychologically) to the faces from which they were modeled (Bailenson, Beall, Blascovich, & Rex, 2004). Figure 3 shows a sample figure with the faces of the nine participants. The only behaviors utilized by avatars were head movements, speech, mouth movements, and blinking (according to a random algorithm based on human blink rates).

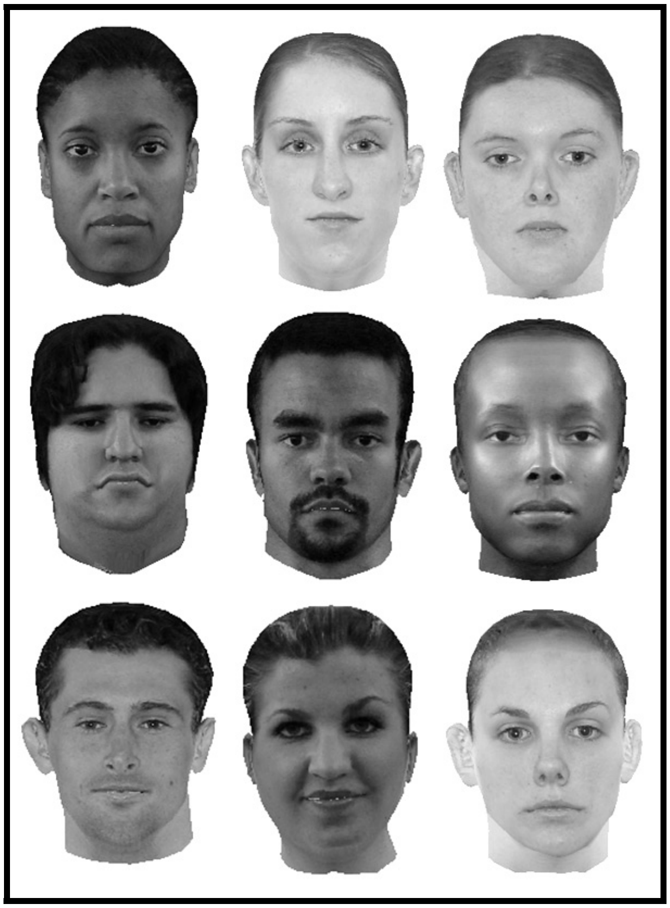


Figure 3: Models used as avatars for our nine participants in this study.

Results

Task Performance

Operationalization of dependent measures. For each of our three types of tasks, we calculated a compound measure of performance based on a combination of quantity, quality, and speed of completion. While we also looked at these component factors independently, the pattern of data from all of the component measures was quite similar to the pattern of data from the compound measures. Consequently, we only report inferential statistics on the compound measures, though means and standard deviations from all component measures are presented in Table 1.

For the 20 Questions task, we computed two component measures. The first was quantity, simply the total number of questions that were used in the three trials. This measure was normalized from 0 to 100, where high scores meant fewer questions were used. The second measure was time, which was the total time used in the three trials. This measure was also normalized from 0 to 100, where a high score meant less time was used. In other words, to get a score of 100, a group on a particular trial would have the fastest score of all trials from all groups. These two measures had a correlation of .70 and were then added together and normalized from 0 to 100 into a single compound measure. In other words, quantity and time were given equal weight in the total normalized compound performance score. This total is not equal to the raw mean of the two scores because we have normalized the sum to allow direct comparisons with other measures.

For the Reverse Association Tasks, we also computed two component measures. The first was quantity – the total number of triads generated across the three trials of the task. This measure was normalized from 0 to 100, where a high score meant more triads were generated. The second measure was quality. Every triad was coded by two raters for quality using the scale provided by Hoyt & Blascovich (2003), and the inter-rater correlation was .58. The sum for the ratings of all triads in each trial was normalized from 0 to 100. These two scores had a correlation of .54 and were then added together and normalized from 0 to 100. In other words, quantity and quality were given equal weight in the total normalized compound performance score.

Finally, for the Insight Problems, we again calculated two measures. The first measure was time used for each trial. This measure was normalized from 0 to 100, where a high score meant less time was used. The second measure was quality. Every answer was rated a four point scale (no answer, poor, unsatisfactory, satisfactory) by two raters. The inter-rater correlation was .92. The rating was then normalized from 0 to 100. These two scores had a correlation of .42. These two scores were then added together and normalized from 0 to 100. In other words, time and quality of answer were given equal weight in the total normalized score.

Data Analysis

Using trial as our random factor and mean performance scores of the triad as the dependent variable, we ran an ANOVA with two repeated measures factors: Task Type (20 questions, RAT, or Insight problems) and condition (face similarity, mimic, normal). Furthermore, we used trial day, the number of days since the first trial for each trial, as a linear covariate. We used trial day (as opposed to trial number) to account for differential amounts of time before certain trials which may have affected participants' familiarity with the equipment. Table 1 shows the means and standard deviations by condition.

Table 1. Means and standard deviations of self-report measures and task performance by condition

	Trial 1			Average Over 15 Trials		
	Face similarity	Mimic	Normal	Face similarity	Mimic	Normal
Copresence	.2 (1.22)	1.2 (.92)	1.2 (.72)	.44 (.59)	.62 (.47)	.96 (.39)
Presence	.22 (.19)	.44 (.84)	.89 (1.35)	-.01 (.32)	-.14 (.40)	.06 (.46)
Entitativity	-.16 (.56)	-.36 (1.76)	-.21 (.06)	.50 (.52)	.53 (.55)	.57 (.55)
Simulator Sickness	1.17 (.00)	1.35 (.22)	1.19 (.10)	1.17 (.08)	1.17 (.10)	1.18 (.06)
20Q Time	54.75	100.00	49.16	35.75 (22.23)	61.97 (31.73)	48.38 (27.07)
20Q Quantity	88.00	44.00	48.80	61.6 (27.41)	39.47 (32.17)	46.99 (16.46)
20Q Total	64.65	18.86	51.52	64.98 (22.60)	39.87 (30.21)	51.22 (21.58)
RAT Quantity	0	25.68	31.69	41.86 (27.32)	46.38 (27.68)	47.32 (30.46)
RAT Quality	87.22	83.78	60.81	50.32 (26.45)	54.89 (26.16)	51.49 (25.79)

RAT Total	87.70	84.24	61.14	50.60 (26.60)	55.19 (26.30)	51.77 (25.94)
IP Time	59.97	54.39	7.94	35.70 (25.96)	32.05 (28.13)	41.49 (32.47)
IP Quality	50.00	0.00	100.00	73.33 (30.57)	75.00 (35.36)	70.00 (35.61)
IP Total	48.93	82.87	49.75	49.30 (19.87)	50.49 (26.26)	47.85 (14.65)

There was no main effect of Condition (*F*[2,26] = 0.33, *p* = .72, Eta-Squared = .01) or Task Type (*F*[2,26]=1.90, *p* = .17, Eta-Squared=.12), however, there was an interaction effect between Condition and Task Type (*F*[4,52]=3.65, *p* = .01, Eta-Squared = .19), as Figure 4 depicts. A Tukey HSD post-hoc test revealed that performance in 20 Questions in the face similarity condition was significantly better than performance in 20 Questions in the mimic condition (p<.05) and the normal condition (p<.06). No other pair-wise comparisons were significant.

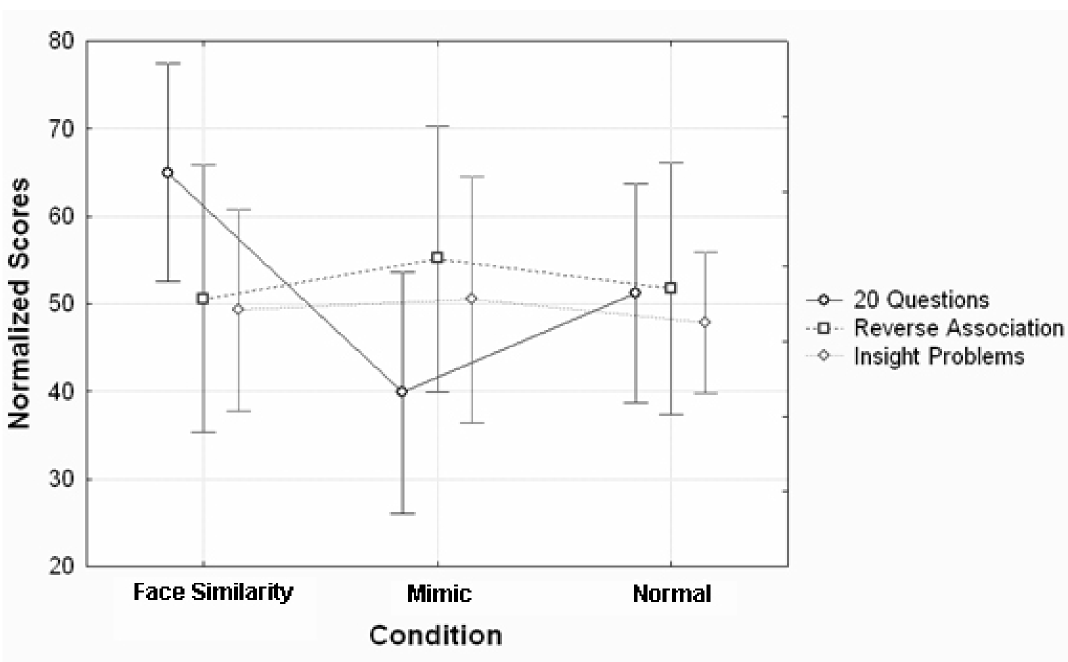


Figure 4: Normalized task performance scores by condition. Error bars denote 95% Confidence Intervals.

Self Report Tasks

There were questionnaires administered at the end of each session: Copresence, Presence, and Entitativity. The reliability for Copresence (Cronbach's alpha = .85) and Entitativity (Cronbach's alpha = .94) were both high. For Presence, Cronbach's alpha was low (below .60) before removing question 2 listed in the Appendix, after which it improved to .75.

Using trial as our random factor and mean ratings of the triad as the dependent variable, we ran an ANOVA with two repeated measures factors: ratings type (Entitativity, Presences, and Copresence) and condition (face similarity, mimic, normal). Furthermore, we used trial day (the number of days since the first trial for each trial) as a linear covariate. Table 1 shows the means and standard deviations by condition.

There was no significant effect of condition, $F(2,52) = 1.18$, $p < .32$, partial Eta-Squared = .08. There was a main effect of ratings type, $F(2,52) = 11.65$, $p < .001$, partial Eta-Squared = .47. Post-Hoc Tukey's HSD tests with an alpha of .05 indicate that both Entitativity and Copresence ratings were significantly higher than Presence ratings but not different from one another. Moreover, as Figure 5 shows, there was a significant interaction between trial day and condition, $F(2,52) = 7.93$, $p < .003$, partial Eta-Squared = .38. Entitativity correlated positively with trial day, $r(15) = .88$, $p < .001$. On the other hand, the correlations of trial day and copresence as well as trial day and presence were negative, but nonsignificant ($r(15) = -.30$, $p < .28$ and $r(15) = -.09$, $p < .74$, respectively).

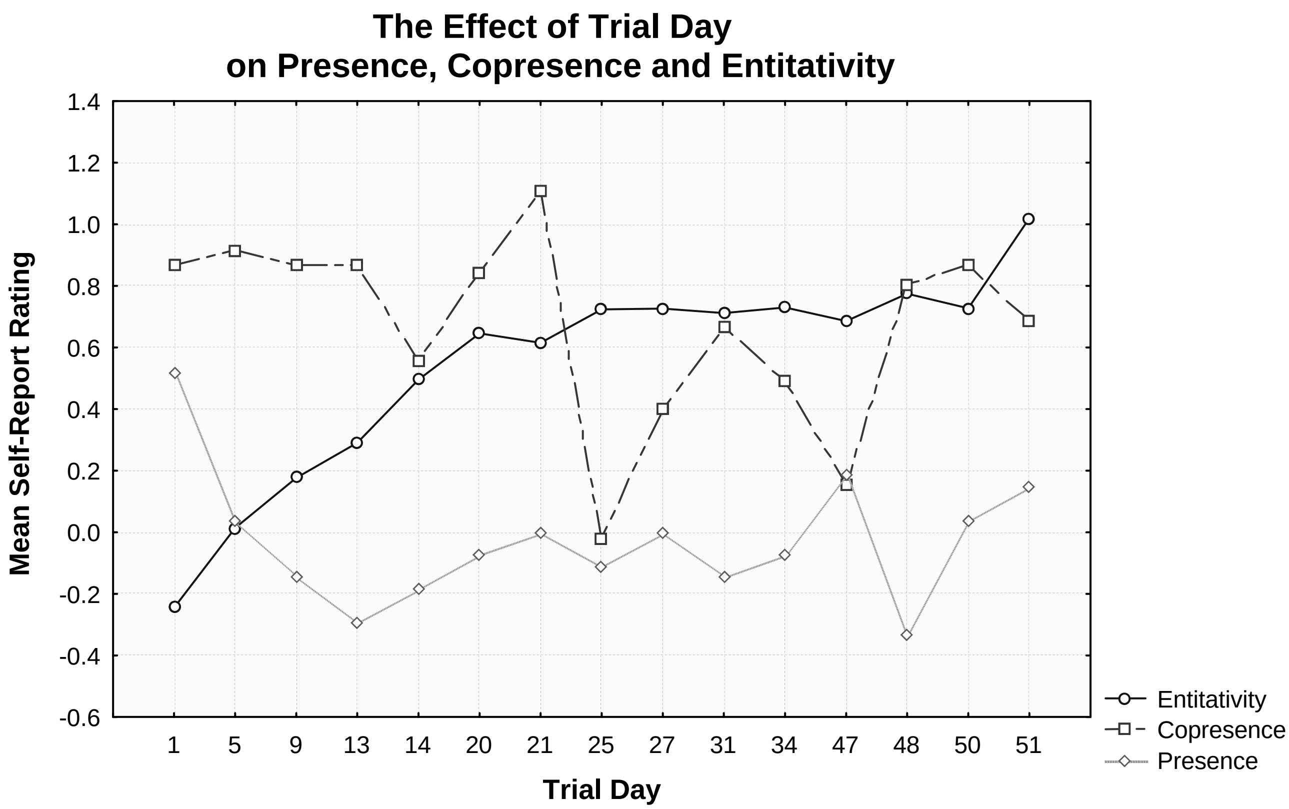


Figure 5: Mean Entitativity, Copresence, and Presence Ratings by Trial Day.

One interesting trend seen in Figure 5 is that self-reported presence ratings become lower over time. While we could not perform any inferential statistics comparing one trial to another with such a small number of participants, this trend is notable and definitely should be examined further in future research.

Head Movements

We next examined the head movements of the participants in the conference. For each participant in each session, we computed a measure of *visual inattention*, the percentage of time that neither of the other two interactants appeared in their field of view. In other words, if the participant was not looking at the head of either of his or her conversational partners, they did not appear in the HMD’s field of view. We used a conservative measure of inattention; if a single pixel of the other avatars appeared in a given participant's field of view it was scored as attention. We ran an analysis of variance with TSI condition as a within-trial independent variable, trial day as a covariate, and visual inattention as a dependent variable. There was no significant effect of condition, $F(2,26) = .73$, $p < .49$, partial Eta-Squared = .05. However, as Figure 6 demonstrates there was an extremely large effect of trial day, $F(1,13) = 30.75$, $p < .0001$, partial Eta-Squared = .70. The more experience participants had with CVEs, the less they looked at one another's faces.

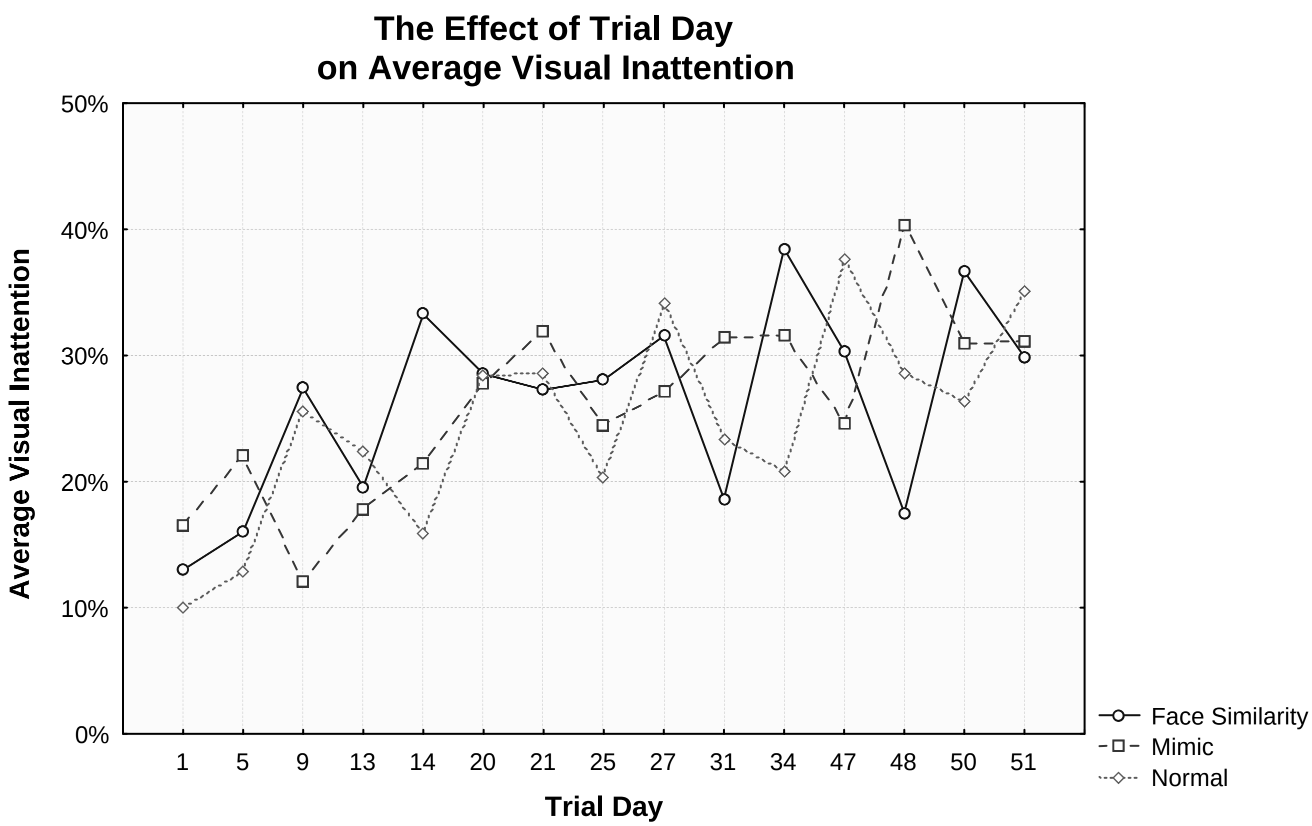


Figure 6: Average Visual Inattention by Condition.

There was no significant interaction between trial number and condition, $F(2,12) = .52$, $p < .60$, partial Eta-Squared = .04.

Simulator Sickness

We took the mean simulator sickness questionnaire score for each trial from all groups. Higher scores indicate more simulators sickness. Figure 7 demonstrates the means for the three conditions. Overall the level of reported sickness was quite low; fewer than five percent of all responses indicated anything but no symptoms or mild symptoms.

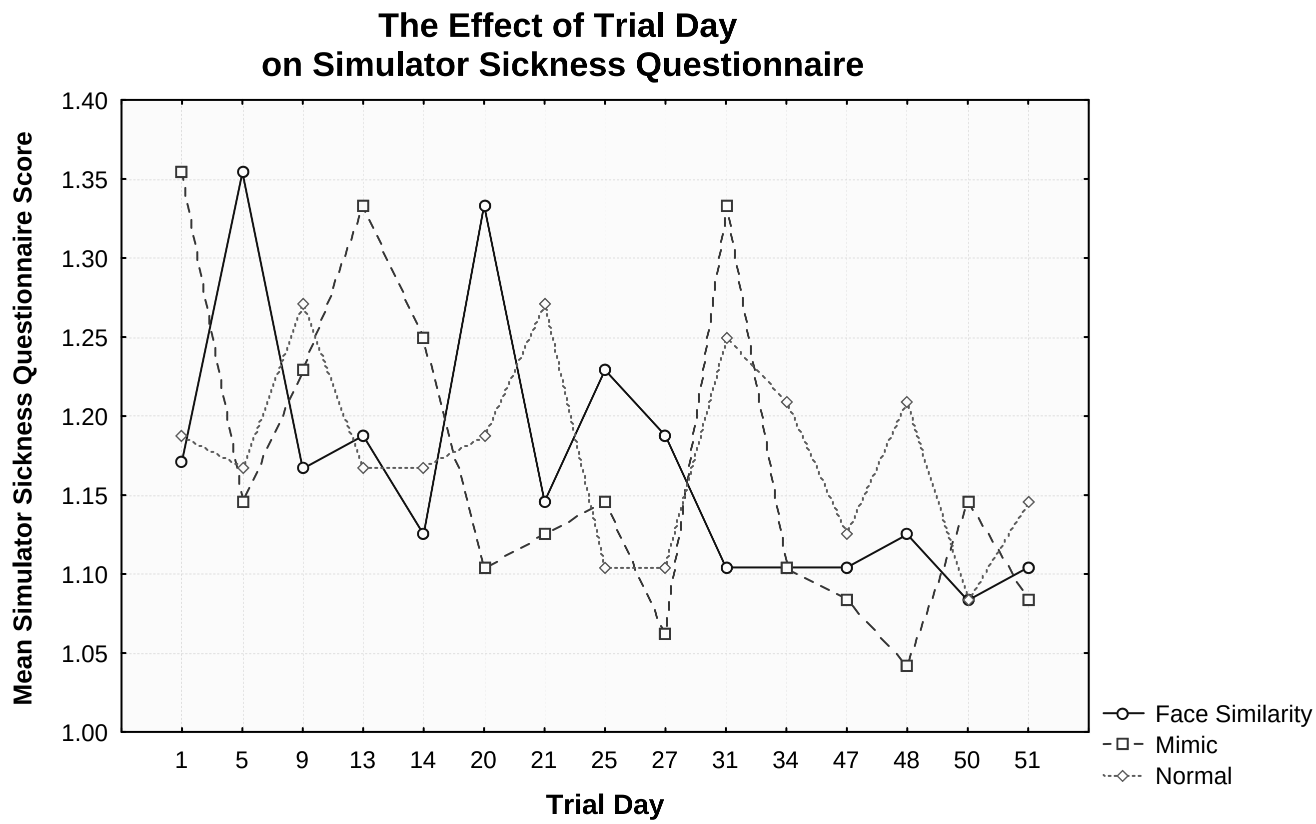


Figure 7: Mean SSQ scores for the three conditions by Trial Day.

We then ran an analysis of variance with TSI condition as a within-trial independent variable, trial day as a covariate, and simulator sickness as a dependent variable. There was no significant effect of condition, $F(2,26) = .57$, $p < .57$, partial Eta-Squared = .04. However, as Figure 7 demonstrates there was an extremely large effect of trial number, $F(1,13) = 35.50$, $p < .0001$, partial Eta-Squared = .73. The more experience participants had with CVEs, the less they reported simulator sickness.

General Discussion

The goal of this study was two fold. First, we implemented two TSI conditions in order to understand how behavioral similarity and visual similarity impacted task performance, presence and copresence. And second, we wanted to explore how task performance, head movements, presence, copresence, simulator sickness, and Entitativity in CVEs changed over time as participants adjusted to the novelty of the technology.

Task performance overall did not vary as a function of the condition (normal, face similarity, mimic), however there was an interaction effect between condition and task type that was driven by the 20 Questions task, in which participants performed best in the face similarity condition. In other words, participants solved the problems using fewer questions in this task when each of them saw their own face on their two conversational partners. This effect is consistent with previous work that demonstrates an increase in task performance when groups are organized to reflect demographic similarity (Kirkman et al., 2004). This effect did not occur with the other tasks; future research should examine the reasons why facial similarity improves performance in some areas but not in others.

One of the most surprising results of this study is the fact that the conversations functioned more or less normally under the various TSI conditions. In the face similarity condition, participants spent a majority of the time looking at other people wearing their exact face—even on the trial with the highest degree of visual inattention, participants looked at one another for approximately 60 percent of the time. Despite post-experimental interviews that indicated that participants found the experience of other people wearing their face to be strange and uncomfortable, participants still looked at one another and actually performed some tasks more efficiently in that condition. Even more striking is the fact that people looked at one another consistently during the mimic condition. For example, Person A paid attention to the head movements of B and C, despite the fact that A clearly was aware that the movements of B and C's avatars were actually his or her own movements, and had nothing to do with the actual behaviors of B and C. Previous research has demonstrated the power of implicit TSI when participants are unaware (Bailenson & Yee, 2005; Beall et al. 2003). The current findings demonstrate that TSI, when made explicit, does not force interactants to behave in a constrained manner. At the beginning of the experiment, we explicitly told participants they were being mimicked during some of the trials. However, during post-experiment interviews, four of the nine participants reported that they had completely forgotten the manipulation.

As the longitudinal study progressed, entitativity increased linearly with the amount of time participants interacted with one another. The fact that the group became more cohesive over time is not surprising. However, it is quite encouraging that this increase in entitativity, which is characteristic of face to face groups, (Lickel et al., 2000) also occurs in CVEs. Furthermore, groups showed high entitativity over time even when their behaviors were altered with TSI. A serious criticism of TSI theory is that when veridical behaviors of group members are replaced by transformed ones, the interactional synchrony (Kendon, 1977) will be lost. In other words, because there may be mismatches between types of behaviors (i.e. verbal behaviors may not match nonverbal behaviors) the intricate structure that governs normal face to face interaction will be lost. The fact that the groups formed extremely high levels of cohesion, even in the presence of TSI, demonstrates the resiliency of people to gain symbolic meaning from even the starkest degree of social cues.

Presence and copresence did not vary significantly over time in the self-report measures, however the head movement data showed that participants looked at each other less over time. Towards the last few sessions, participants in all three TSI conditions were looking at one another for only approximately 60 percent of the time during the interaction, compared to approximately 85 percent for the first few trials. While visual contact should have an impact on measures of copresence, we did not find any differences in self-report of copresence over time. Thus, these findings suggest that important behavioral changes were not captured by the self-report measures, which may be too rough of a measurement tool to capture important psychological aspects of CVE use, as previous work has also suggested (Bailenson et al., in press; Slater, 2004). Finally, self-reported simulator sickness declined as participants became more familiar with the virtual reality equipment. Previous work that draws conclusions about the nature of simulator sickness, without examining their systems over time, should be wary of their results.

There are a number of limitations of the current study. First of all, due to the small sample size of participants, we did not have enough statistical power to examine the data as thoroughly as we would have liked to. Furthermore, participants had met face-to-face before beginning the study (and often saw one another on the way to the laboratory) and this level of familiarity most likely added noise to our entitativity and copresence measures. Previous work has demonstrated that meeting face-to-face before meeting in a computer-mediated setting causes the group to be more cohesive than mediated meetings without prior face-to-face interaction (Rocco, 1998). Future work should attempt to study longitudinal development of these constructs in CVEs that are physically remote in order to avoid this contamination to the data. Moreover, the lack of a difference in TSI conditions on some of our dependent variable may be explained by a lack of motivation of participants to perform the specific experimental tasks. In future work, increasing participant motivation with monetary incentives may produce different results. Finally, it could be the case that by presenting conditions in continuous blocks (as opposed to randomly dispersing them) would have produced more pronounced differences by allowing subjects to familiarize themselves with the communication paradigm. Similarly, having a control condition that was either voice only or face-to-face would allow more direct comparisons between CVEs and other communication media directly. Future work should examine this methodology.

These findings on head movement changes over time have larger implications on CVE usage. They suggest that over time, the participants paid much less attention to the visual resources in the environment and relied more heavily on auditory resources. Thus, ironically, it appears that over time the participants occasionally used the CVE more as a telephone conference than a visual conference. While most likely this trend is in part driven by our specific CVE implementation (i.e., the small number of actual user behaviors tracked and rendered which would provide useful social cues as well as the limited HMD field of view, weight, resolution, etc.), the fact remains that people relied less and less on visual avatar representation as they became more familiar with the CVE technology. This behavioral data is mirrored by the fact that subjective ratings of presence were much higher in the initial two sessions than in all other sessions, although due to our low number of participants it was not possible to test statistical significance of this trend.

The findings also make clear that important behavioral changes do occur over time in CVE interactions and that it is naïve to assume that the novelty of the technology and equipment has no impact on the interaction among first-time users. This is particularly true for almost all CVE studies that have been run using first-time users. The technology is so novel and difficult to use that the manner in which first-time users employ these systems are vastly different from users who have more experience with the technology.

Footnotes

[1]Due to computer error, on two of the fifteen trials, experimental conditions were received by more than one group, as Appendix A demonstrates. For purposes of inferential statistical tests, we switched the data/conditions of group B for trial 9 and for trial 11. This switch preserved the integrity of the data files better than replacing missing values with the mean of that condition would have.

Author Notes

The author's would like to thank Jakki Oni Bailey, Emily Boyd, Claire Carlson, KC Corkery, Gerren Crochet, Melanie Curtain, Alexandra Davies, Eric Ford, Allison Lee, Malia Skaer, Holly Thorsen, and Sam Warburg for the assistance during this project. Furthermore, we would like to thank Robby Ratan for helpful comments on the paper.

Appendix A: Trial Schedule

Trial	Task1	Task2	Task3	Group A (days since last trial)	Group B (days since last trial)	Group C (days since last trial)
1	20 Questions	RAT	Insight	face sim	normal	mimic
2	20 Questions	RAT	Insight	normal (5)	mimic (5)	face sim (5)
3	RAT	Insight	20 Questions	mimic (7)	face sim (2)	normal (2)
4	Insight	20 Questions	RAT	face sim (2)	normal (5)	mimic (5)
5	RAT	20 Questions	Insight	normal (0)	mimic (1)	face sim (2)
6	20 Questions	Insight	RAT	mimic (5)	face sim (8)	normal (5)
7	Insight	RAT	20 Questions	face sim (2)	normal (1)	mimic (2)
8	Insight	RAT	20 Questions	normal (5)	mimic (4)	face sim (5)
9	20 Questions	Insight	RAT	mimic (2)	mimic (2)	normal (2)
10	RAT	20 Questions	Insight	face sim (0)	normal (6)	mimic (5)
11	Insight	20 Questions	RAT	normal (2)	face sim (1)	face sim (7)
12	RAT	Insight	20 Questions	mimic (17)	face sim (21)	normal (2)
13	20 Questions	RAT	Insight	face sim (0)	normal (0)	mimic (1)
14	RAT	20 Questions	Insight	normal (2)	mimic (1)	face sim (4)
15	20 Questions	Insight	RAT	mimic (0)	face sim (0)	normal (2)

Note: Zero days since the last trial indicates that the trial was on the same day.

Appendix B: Task Stimuli

20 Questions and RAT Words

	20 Q Word 1	20 Q Word 2	20 Q Word 3	RAT Word 1	RAT Word 2	RAT Word 3
Trial 1	Boat	Monkey	Happiness	Deep	Bar	Jack
Trial 2	Pineapple	Communism	Bookshelf	Blue	Blast	White
Trial 3	Astronomy	Soap	Butterfly	Salt	Pipe	Music
Trial 4	Telephone	Cactus	Health	Party	Gift	Cabbage
Trial 5	Snake	Embarrassment	Scissors	Candle	Gin	Thief
Trial 6	Fire	Toothbrush	Dolphin	Quiet	Paint	Sea
Trial 7	Wallet	Grass	Gender	City	Beer	Dull
Trial 8	Earthworm	Poetry	Ring	Card	High	Cold
Trial 9	Anger	Umbrella	Carrot	Needle	Note	Sack
Trial 10	Bread	Spider	Music	Silver	Cheese	Eagle
Trial 11	Pigeon	Intelligence	Book	Cookie	Glass	Street
Trial 12	Color	Milk	Wolf	Memory	Water	Tape
Trial 13	Salt	Squirrel	Paris	Pool	Paper	Book
Trial 14	Penguin	Christmas	Fork	Bath	Smooth	Table
Trial 15	Beauty	Bicycle	Banana	Spider	Head	Top

Insight Problems

Problem 1 – Radiation

Question: A patient with a tumor in his stomach will die unless he is treated, but he cannot be operated on. There is a new laser treatment available. The ray of laser can be adjusted in intensity. Unfortunately, the intensity necessary to destroy the tumor will also destroy healthy tissue the ray of laser passes through, and at intensities that will not hurt healthy tissue, the ray of laser will not affect the tumor. How can the patient be saved using this laser treatment?

Answer: Use multiple lasers directed at the same spot, so the intensity is only strong enough in the spot where all lasers meet.

Problem 2 – Water Lilies

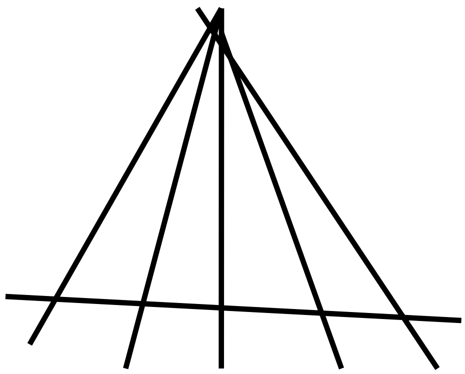
Question: The water lilies in a pond double every 24 hours. If the pond is filled with water lilies in 30 days, on what day was the pond half full?

Answer: On day 29, because on the day before, the pond would have been half full.

Problem 3 – Matchsticks

Question: What is the smallest number of matchsticks you need to form 10 triangles? You may not break the matchsticks.

Answer: 6



Problem 4 – Animal Pens

Question: How would you put 27 animals in 4 pens such that there are an odd number of animals in each pen?

Answer: Put all animals in one pen, then have 3 concentric pens around it.

Problem 5 – Oil and Vinegar

Question: A small bowl of oil and a small bowl of vinegar are placed side by side. You take a spoonful of the oil and stir it casually into the vinegar. You then take a spoonful of this mixture and put it back in the bowl of oil. Which of the two bowls is more contaminated?

Answer: They are equally contaminated.

Problem 6 – Knights and Knaves

Question: On a particular island, there are only knights and knaves.

1) Knights always tell the truth
2) Knaves always lie

You meet Alice, Bob, and Rex. Alice says Rex is a knave. Rex says Bob is a knave. Bob says either he is a knight or Alice is a knight. What are Alice, Bob and Rex.

Answer: Alice is a knave. Rex is a knight. Bob is a knave.

Problem 7 – 20 Quarters

Question: You are wearing blindfolds and gloves. On a table in front of you are 200 quarters. You know that 20 of them are heads, while the rest are tails. Your task is to split the quarters into two piles where both piles have the same number of heads. How do you accomplish this?

Answer: Take any 20 quarters. Move them into a separate pile. Flip them over.

Problem 8 – 3 Light Bulbs

Questions: You are in a room with 3 light switches, each of which controls 1 of 3 light bulbs in the next room which you cannot look into. You are allowed to enter the room with the light bulb only once. How can you determine which switch is connected to each of the 3 bulbs?

Answer: Turn one switch on for 5 minutes. Turn it off. Turn another switch on. Go into room. There will be one bulb that is on, one bulb that is off but warm to touch, and one bulb that is off and cold.

Problem 9 – Cork and Bottle

Question: You put a coin into a wine bottle and then fit a cork snugly into the bottle's opening. How can someone remove the coin without breaking the bottle or removing the cork?

Answer: Push cork into bottle. Pour coin out.

Problem 10 – 4 Trees

Question: A landscape designer is asked to place four trees such that they are equidistant from each other. How does he accomplish this?

Answer: Has to be in pyramid shape. So must be planted on a hill or some similar terrain.

Problem 11 – River Crossing

Question: Two men arrive at a river crossing at the same time. There is only one boat and the boat can only carry one person at a time. Neither of them can swim and there is no other way to cross the river, yet they both manage to cross the river. How did they accomplish this?

Answer: They start off on different sides of the river.

Problem 12 – Rope Burning

Question: You have two badly-woven topes with fluctuating widths, but which you know take one hour each to burn completely. But because of their fluctuating widths, thicker parts take longer to burn, so half of each rope may take longer or shorter than 30 minutes to burn. Using these two ropes, how can you time 45 minutes?

Answer: Burn one rope from both ends, and the other from one end. When the first rope is done, 30 minutes will have passed, and 30 minutes will be left on the second rope. At that point, light the second rope on its other end. When the second rope burns out, 15 minutes will have passed.

Problem 13 – Locks

Question: Andy and Brian are staying in different rooms in a hotel. Andy needs to give a golden pendant to Brian, but spies are in the hotel so neither Andy nor Brian can leave their rooms. The only way they can transfer items is through bellhops, but the bellhops will steal anything that is not secured in a portable safe. Andy and Brian each have a safe with a large clasp that can be secured with a padlock. Both Andy and Brian have their own padlock with a corresponding key. So 1 pendant, 2 safes, 2 padlocks, and 2 keys, But anything that is not locked in a safe will be stolen by the bellhops. How can Andy get the pendant to Brian?

Answer: Andy puts pendant in safe and padlocks it and sends it to Brian. Brian double-padlocks the safe and sends it back to Andy. Andy removes his lock and sends it back to Brian. Brian removes his padlock.

Problem 14 – Gold Coin

Question: There are 10 machines that produce gold coins that each weigh 2 pounds. You know that one of these machines is defective and making coins that are always 1 pound lighter than they should be. You are given an electric scale that gives you the exact weight of what is on it. How do you figure out which machine is defective with just one weighing?

Answer: Take 1 gold coin from 1st machine, 2 coins from the 2nd machine, and so on. Weigh them. The number of pounds missing would refer to the defective machine.

Problem 15 – Letter Sequence

Question: What is the next letter in the following sequence?

O T T F F S S E N ?

Answer: One Two Three Four Five Six Seven Eight Nine Ten = T

Appendix C: Questions from self report measures

Copresence (7 point response scale with endpoints labeled as "agree" or "disagree")

I perceived that I was in the presence of other people in the virtual room with me.
I felt that the people in the virtual room were watching me and were aware of my presence.
The thought that they were not real people crossed my mind often in the virtual room (r).
The people in the virtual room appeared to be sentient (conscious and alive) to me.
I perceived the people as being only a computerized image, not as real people (r).

Presence (7 point response scale with endpoints labeled as "agree" or "disagree")

I forgot about my immediate physical surroundings (i.e., the lab room) when I was in the virtual conference.
I paid more attention to my own thoughts (e.g., personal preoccupations, daydreams, etc.) than what was going on in the virtual conference (r).
I did not want to reach out and touch things in the virtual conference (r).
I felt like I was in a psychology laboratory rather than a virtual conference (r).

Entitativity (7 point response scale with endpoints labeled with content specific endpoints)

Please indicate to what extent you would use the term "WE" to characterize you and the other two people in this group by circling the appropriate number.
Please indicate to what extent you and the people in this group are currently "connected".
Please indicate how well you know the other two people in this group by circling the appropriate number.
Please indicate how much you like the other people in this group by circling the appropriate number.
Please indicate how much you consider the three people in your group to be 'a group'
Please indicate how connected to one another you consider the people in your group to be.
How related to one another are the members of your group?
Some sets of people are very much a "group"; when thought of, looked at, or described they are treated as a single entity, as "one thing." Other sets of people, on the other hand, are not much of a "group"; when thought of, looked at, or described they are treated as a number of individuals distinct from one another. With that in mind, you would describe the caste as one group or not at all one thing or very much one thing?
When we think about certain things in the world, we sometimes think about the components that make them up. Other times, we think about the things as single units. With this distinction in mind, when you think about the caste, you think of its components or a single unit?
When we consider a set of people, we sometimes think about them "as a whole", and consider the characteristics of the people in general. Other times, we think about the

set of people as a number of individuals. If you were thinking about these people, would you think about them as a number of individuals or as a whole?

Simulator Sickness Questionnaire (4 point response scale with endpoints labeled as "No Symptoms" or "Severe")

Please indicate the extent to which you are presently experiencing any virtual environment aftereffects RIGHT NOW.

General Physical Discomfort, Fatigue, Headache, Eyestrain, Difficulty Focusing, Increased/Decreased Salivation, Sweating, Nausea, Difficulty concentrating, Fullness of Head (like "medicine head"), Blurred Vision, Dizzy (eyes open), Dizzy (eyes closed), Vertigo (feeling like you or the world is spinning, Stomach Awareness, Burping.

References

Bailenson, J. N., & Beall, A. (2005, in press). Transformed social interaction: Exploring the digital plasticity of avatars. In R. Schroeder & A. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments*: Springer-Verlag.

Bailenson, J. N., Beall, A. C., & Blascovich, J. (2002). Mutual gaze and task performance in shared virtual environments. *Journal of Visualization and Computer Animation, 13*, 313-320.

Bailenson, J. N., Beall, A. C., Blascovich, J., & Rex, C. (2004). Examining virtual busts: Are photogrammetrically-generated head models effective for person identification? *PRESENCE: Teleoperators and Virtual Environments, 13(4) 416-437*.

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., 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., Garland, P., Iyengar, S., & Yee, N. (in press). The effect of implicit similarity in political communication. *Political Psychology*.

Bailenson, J. N., Swinth, K. R., Hoyt, C. L., Persky, S., Dimov, A., & Blascovich, J. (in press). 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: Teleoperators and Virtual Environments*.

Bailenson, J. N., & Yee, N. (in press). Digital chameleons: Automatic assimilation of nonverbal behaviors in immersive virtual environments. *Psychological Science*.

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

Benford, S., Bowers, J., Fahlen, L., Greenhalgh, C., Mariani, J., & Rodden, T. (1995). Networked virtual reality and cooperative work. *PRESENCE: Teleoperators and Virtual Environments, 4*, 242-249.

Benford, S., Bowers, J., Fahlen, L., Greenhalgh, C., & Snowdon, D. (1995). User embodiment in collaborative virtual environments. Paper presented at the *CHI '95 Electronic Proceedings*.

Berscheid, E., & Walster, E. (1974). Physical attractiveness. *Advances in Experimental Social Psychology, 7*, 157-215.

Byrne, D. (1971). *The attraction paradigm*: New York, NY: Academic Press.

Casanueva, J., & Blake, E. H. (2000). The effects of avatars on co-presence in a collaborative virtual environment. Paper presented at the *EGVE'00 - 6th Eurographics Workshop on Virtual Environments*, June 2000.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality & Social Psychology, 76*(6), 893-910.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces. Paper presented at the *CSCW 92 Proceedings*.

Dovidio, J. F., Gaertner, S. L., Anastasio, P. A., & Sanitioso, R. (1992). Cognitive and motivational bases of bias. In S. B. Knouse, P. Rosenfeld & A. L. Culbertson (Eds.), *Hispanics in the workplace* (pp. 75-106): Newbury Park, CA: Sage.

Garau, M., Slater, M., Vinayagamoorhty, V., Brogni, A., Steed, A., and Sasse, M.A. (2003). The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. *Proceedings of the SIG-CHI*, 529 - 536.

Hoyt, C. L., & Blascovich, J. (2003). Transformational and transactional leadership in virtual and physical environments. *Small Group Research, 34*(6), 678-715.

Huesmann, L. R., Lagerspetz, K., & Eron, L. D. (1984). Intervening variables in the tv violence-aggression relation: Evidence from two countries. *Developmental Psychology, 20*(5), 746-777.

IJsselsteijn, W. A., Freeman, J., Avons, S. E., Davidoff, J., de Ridder, H., & Hamberg, R. (1997). Continuous assessment of presence in stereoscopic displays. *Perception, 26,* 42-43.

Kendon, A. (1977). *Studies in the behavior of social interaction*.Bloomington: IU.

Kennedy, R. S., Lane, Berbaum & M. G. Lilienthal (1993). A simulator sickness questionnaire (SSQ): A new method for quantifying simulator sickness. *International Journal of Aviation Psychology* 3(3): 203-20.

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2004). The impact of demographic heterogeneity and team leader-team member demographic fit on team empowerment and effectiveness. *Group and Organization Management, 29*(3), 334-368.

Lanier, J. (2001). Virtually There, *Scientific American*, April, 66-75.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology*, 78, 223-246.

Maner, J. K., Luce, C. L., Neuberg, S. L., Cialdini, R. B., Brown, S., & Sagarin, B. J. (2002). The effects of perspective taking on motivations for helping: Still no evidence for altruism. *Personality and Social Psychology Bulletin, 28*, 1601-1610.

Normand, V., Babski, C., Benford, S., Bullock, A., Carion, S., Chrysanthou, Y., et al. (1999). The coven project: Exploring applicative, technical, and usage dimensions of collaborative virtual environment. *PRESENCE: Teleoperators and Virtual Environments, 8(2),* 1999.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York, NY: Center for the Study of Language and Information Cambridge University Press.

Rocco, E. (1998). Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact, *Proceedings of the Conference on Human Factors in Computing Systems (CHI'98),* 496-502.

Schroeder, R., & Axelsson, A.-S. (2000, May 12 - 13). *Trust in the core: A study of long-term users of activeworlds.* Paper presented at *A Cybercultural Symposium* in Norrkoping, Sweden.

Shanteau, J., & Nagy, G. (1979). Probability of acceptance in dating choice. *Journal of Personality and Social Psychology*, *37*, 522-533.

Slater, M. (in press). How colourful was your day? Why questionnaires cannot assess presence in virtual environments. *PRESENCE: Teleoperators and Virtual Environments*.

Steed, A., Slater, M., Sadagic, A., Bullock, A., & Tromp, J. (1999). Leadership and collaboration in shared virtual environments. *IEEE VR '99*.

Suzuki, N., Takeuchi, Y., Ishii, K., & Okada, M. (2003). Effects of echoic mimicry using hummed sounds on human-computer interaction. *Speech Communication, 40*(4), 559-573.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the stanford revision and extension of the binet-simon intelligence scale*: Boston: Houghton Mifflin.

van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393-398.

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. J. Sternberg & J. E. Davidson (Eds.), (Vol. The Nature of Insight, pp. 157-196): Cambridge: MIT Press.

Yee, N. (in press). The Demographics, Motivations and Derived Experiences of Users of Massively-Multiuser Online Graphical Environments. *PRESENCE: Teleoperators and Virtual Environments*.

Zhang, X., & Furnas, G. (2002). *Social interactions in multiscale CVEs.* Paper presented at the *Proceedings of ACM Conference on Collaborative Virtual Environments*, 2002.