Communication Behavior in Embodied Virtual Reality

Harrison Jesse Smith[1,2] **and Michael Neff**[1,2]
[1]Oculus Research, Sausalito, CA, USA
[2]University of California, Davis, CA, USA
michael.neff@oculus.com, hjsmith@ucdavis.edu

ABSTRACT
Embodied virtual reality faithfully renders users' movements onto an avatar in a virtual 3D environment, supporting nuanced nonverbal behavior alongside verbal communication. To investigate communication behavior within this medium, we had 30 dyads complete two tasks using a shared visual workspace: negotiating an apartment layout and placing model furniture on an apartment floor plan. Dyads completed both tasks under three different conditions: face-to-face, embodied VR with visible full-body avatars, and no embodiment VR, where the participants shared a virtual space, but had no visible avatars. Both subjective measures of users' experiences and detailed annotations of verbal and nonverbal behavior are used to understand how the media impact communication behavior. Embodied VR provides a high level of social presence with conversation patterns that are very similar to face-to-face interaction. In contrast, providing only the shared environment was generally found to be lonely and appears to lead to degraded communication.

ACM Classification Keywords
H.4.3. Communications Applications: Computer conferencing, teleconferencing, and videoconferencing

Author Keywords
Computer-mediated communication, virtual reality, embodiment, social presence.

INTRODUCTION
Modern communication is frequently mediated by technology. Each technology offers its own set of affordances [12], and yet it is difficult to match the immediacy and richness offered through the multimodality of face-to-face communication – so much so that the Department of Energy estimates that roughly eight percent of US energy is used to support passenger transport to enable face-to-face communication [1]. This work explores how embodied virtual reality (VR) can support communication around a spatial task. Embodied virtual reality means that a person's movements are tracked and then used to drive an avatar in a shared virtual world. Using a head mounted display (HMD), participants view the world through the avatar's eyes, and the avatar's movements reflect those of their own body, effectively embodying them in the virtual world. This technology allows people to interact in a shared, 3D environment and employ both verbal and nonverbal communication. In this work, we use marker-based, optical motion capture to track participants and render their bodies as simple 3D meshes in the environment, with an eyebrow ridge and nose, but no other facial features or facial animation (Fig. 1-C).Such an embodiment supports manual gesture, locomotion and verbal dialog, but limited hand movement and no facial expressions.

To understand how such a representation (embodVR) performs as a communication tool, we compare it to two other conditions. The first is the gold standard: face-to-face communication (F2F). The second is a variant of VR in which both participants can see the same shared environment, but their avatars are not visible to themselves or each other (no_embodVR) (One task provided a representation of participant's own hands to facilitate object interaction). Employing a within-subject design, 30 dyads interacted with each other in each of the three conditions, performing two tasks in each. They were told to role-play being new roommates, and in the first task, they were given a floor plan of their new apartment and had to agree on which rooms should be used for the living room, dining room and each of their bedrooms. In the second task, they had to agree on a furniture arrangement by placing representative furniture on the floor plan. These tasks offer an appropriate test bed for interactions in which people must work with a shared visual representation. The first task does not require manipulation of the environment and the second does.

The technologies were evaluated based on both participants' subjective impressions and a detailed analysis of their actual verbal and nonverbal communication behavior. We expected to see three distinct levels of performance, where F2F performs best, followed by embodVR and then no_embodVR. This echoes earlier work comparing face-to-face, audio/video, and audio-only communication. Instead, we found very similar behavior between F2F and embodVR, but a marked drop off for no_embodVR. Recordings (real and virtual) of the apartment layout task were annotated for both verbal and nonverbal behavior. For most factors, there was no significant difference between F2F and embodVR, with often a significant drop off for no_embodVR. This suggests that participants employed similar communication patterns in F2F and embodied virtual reality. Subjective measures provide insight into the "felt experience" of using the technology. On most, but not all, measures

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173863

of social presence, the same pattern emerged of no significant difference between embodVR and F2F, but a significant drop off for no_embodVR. Much more detail is provided below.

RELATED WORK

Most real-time collaboration mediums can be grouped into three different categories: face-to-face, tools that support audio communication only, such as a telephone, and tools that support audio and visual communication, such as video conferencing (for detailed reviews of computer-mediated communication, please see [16, 47]). Previous works have established a hierarchy in which face-to-face interactions are clearly superior to audio-only for most tasks involving spatial information or negotiation [47]. The role video plays is less clear-cut, however. While providing a visual channel can theoretically aid in mutual understanding (or *conversational grounding*), tools with video feeds often do not perform significantly better than audio-only equivalents [35, 32, 43, 25]. There are a number of fundamental issues preventing video conferencing tools from reaching the effectiveness of face-to-face interactions:

1. Interlocutors connected by video feeds are not co-present: they cannot fully observe their partner (visibility) or their partner's physical context (visual co-presence). Fussell and Setlock argue that there is "clear evidence" that visual co-presence improves task performance [16] and helps support grounding [18]. Visibility allows people's gestures to be seen, which is important for representational gestures, and co-presence allows them to point to features in the environment. Co-presence reduces the need for verbal grounding. In embodied virtual reality, participants are naturally co-present in the same virtual space, without a need to try to engineer these features into a remote video system.

2. Most video feeds are stationary, positioned to provide close-up representations of the remote partner's face or upper torso. Such positioning can hinder movement-centric tasks, and prevent transmission of posture and gesture cues. Most non-stationary video feeds, such as smart phone cameras, are controlled by the remote partner, which can result in disorienting camera movements and sub-optimal camera angles [26]. In embodied virtual reality, participants control their view of the scene by adjusting their gaze.

3. Offsets between screens and cameras make it difficult or impossible to establish mutual gaze. Eye contact is important for conversational management and establishing intimacy between partners, and its absence can reduce the perceived social presence of the communication tool. The difficulty in establish eye contact can disrupt conversational management behaviors [27]. Observing gaze also allows one to infer a partner's focus of attention, which can aid grounding.

The Role of Shared Visual Workspaces in Collaboration Tasks

When considering remote collaboration tools with a visual component, it is helpful to draw distinctions between shared visual information pertaining to the state of a task (shared visual workspace) and visual information depicting the state of the remote partner. Many previous works have focused on shared visual workspaces as a key component of effective remote collaborations [19, 20, 21]. Because the current study focuses on evaluating the impacts of embodiment, all conditions incorporate a shared visual workspace.

When performing complex tasks with many different possible states, a shared visual workspace can help partners synchronize their mental models of the task state and aid in the development of common ground: Gergle et. al. [19] found that, when completing a puzzle, the speed and efficiency with which dyads completed the task increased when a shared workspace was present. Participants were less likely to verbally verify their actions, relying on the visual information to transmit the necessary communicative cues. In a study conducted by Fussell et al. [15], partners completed bicycle repair tasks under various conditions: the task was most efficiently completed in the face-to-face condition, and participants attributed this to the presence of an effective shared visual workspace.

Interestingly, a shared visual workspace does not always result in more efficient communication: in the same study by Fussell et al., the addition of a video feed depicting the task did not show significant advantages over the audio-only condition: participants mentioned that it was difficult to make effective deictic gestures [15]. In related studies, Kraut et. al. [32, 30] performed an experiment where novice/expert pairs of bicycle mechanics interacted via audio and audio/video media. While the presence of a shared visual workspace did assist in enabling shared common ground and more proactive help-giving by the expert, it did not ultimately result in higher quality solutions.

The Role of Visual Behavior in Collaboration Tasks

Video feeds that show a remote partner allow users to communicate with a wide array of nonverbal behaviors (such as gaze, gesture or posture) that can influence characteristics of the interaction; comparisons of audio telephony to face-to-face conversation indicate that people use more words and turns in audio-only conditions [16]. When people can gesture at a work space, they can use deictic utterances ("that", "those", "there", etc.) rather than relying on longer, more descriptive phrases ("the table between the door and the window"). Studying a drawing task, Bly found that gesturing behavior decreased both in count and as a percentage of drawing surface actions as interaction moved from from face-to-face to a video-link to telephone only [10]. Clark and Krych [13] found that the presence of a visual workspace results in many more deictic utterances and gestures. Some experimental work has shown that, if these gestures are appropriately supported, resulting communication is more efficient, task performance increases or users rate the quality of the experience higher [17, 28, 2].

The presence of nonverbal modalities may have additional, subtler effects. Peksa et. al. found that, in VR settings, orienting avatars towards a user results in the user taking significantly more turns in a conversation [41]. Fussell and Setlock [16] mention that conversational turn taking becomes more formal in the absence of visual cues. Bailensen et. al. found that, when virtual humans made eye contact with users, female users tended to stand further away [5].

Evaluating Communication Tools

A useful remote communication tool should be efficient, allowing users to quickly achieve high-quality solutions. Therefore,

two commonly used metrics are the quality of task solutions achieved and their times-to-completion [32, 39, 48]. More nuanced insights can be gained by annotating and analyzing interactions: length, frequency, and initiation of conversational turns [37], gaps between turns [25], gaze patterns [4], and overlapping turns [43]. For example, overly-elaborate, highly redundant messages may indicate that a communication tool does not adequately support backchanneling [29, 31], which can result in decreased mutual understanding between partners. Presence of deixis, shorter spaces between conversational turns, shorter, more frequent conversational turns and the presence of interruptions can all provide important clues about how comfortable users feel within a medium.

More subjective measures are obtained through user surveys. One such measure comes from Social Presence theory, which argues that technologies differ in their abilities to present a sense of other participants' goals and motives [44]. A widely-used test for social presence, semantic differencing [40], asks users to evaluate the communication tool by rating it along multiple pairs of bipolar adjectives. It has been shown to be sensitive to differences between 2D video, 3D video, and face-to-face interactions[23]. It has also been used to distinguish between video conferencing tools that do and do not support spatiality [24, 22]. Networked minds [9, 8, 7] is an alternative survey approach focused on whether the user experienced presence; both surveys were administered in the current work.

METHODS

We designed a study to evaluate communication differences between face-to-face interactions (F2F), VR with a motion-capture tracked avatar providing an embodied representation of the users (embodVR) and VR with no avatars (no_embodVR). All conditions included a visually shared work space.

Participants

A total of 60 subjects (30 male, 30 female) were recruited through a professional agency, along with a backup roster of friends, remote coworkers, and neighbors. In all cases, care was taken to make sure participant pairs were strangers prior to beginning the experiment. FannoParticipants were paired into 30 same-gender dyads to limit the number of combinations and remove a potentially confounding factor of strangers being less comfortable with the roommate scenario when dealing with an opposite gender stranger. Participants were aged 18-56 (M=36.5, SD=10.0). The experiment took approximately 3 hours to complete and participants were compensated with gift cards. IRB approval was obtained ahead of the experiment.

Study Design

The experiment employed a 1x3 design in which the factor, communication medium, was (1) face-to-face, (2) virtual reality with a shared workspace, audio channel, and visible, fully-embodied, motion-capture driven avatars, or (3) virtual reality with a shared workspace and audio channel, but no avatars. To control for variations within participant pairs, we employed a within-subjects design where each dyad performed the tasks under each of the three conditions. In order to prevent subjects from reusing previous decisions in subsequent conditions, three different floor plans were utilized. The order in which the floor plans were utilized was constant across all dyads, confounding its effect with that of factor ordering. The impacts of both were minimized by using a fully-counterbalanced design (five repetitions of each of the six possible factor orderings). Post-experiment tests showed that order never had a significant impact on any of the factors examined.

Because the effect of a communication medium can depend upon the type of task being conducted, our participants completed two distinct tasks under each condition: first a negotiation task, then a consensus-building workspace-manipulation task. These tasks are both social and relied on a shared visual workspace, allowing comparison of the role of the shared workspace with the visual presence of the interlocutor. We chose tasks that were familiar, but allowed detailed discussion.

Task 1: Negotiating Room Uses

Participants were instructed to role play a pair of roommates moving into a new apartment. They were given a miniature version of the apartment floor plan, which contained a labeled bathroom and kitchen, a lake, a road, and four unassigned rooms, labeled 'A' through 'D'. These labels facilitated easy verbal references. Participants then decided which of these rooms would be each participant's bedroom, the dining room, and the living room. The participants were given non-compatible preferences (both wanted the same bedroom, and different rooms to be the living room) and told to role play as they saw fit in justifying their preferences. Participants were given five minutes for the task and rooms were assigned by the researcher if consensus was not reached.

Task 2: Building Consensus for Furniture Placement

In the second task, participants placed furniture inside the apartment for which they had just assigned rooms. To foster inter-partner communication during the process, participants were asked to take turns placing furniture while adhering to a specific protocol. For each turn, one participant would select a piece of furniture and suggest two different locations for it inside the apartment, justifying each location. Their partner would then suggest a third option and justify it. Then, both partners would discuss the options and select one together. After this, the participants would switch roles for the next turn. Participants completed this task for for ten minutes.

Procedure

Upon arriving at the testing facility, participants completed a consent form and a short demographic survey. During this period, the researcher confirmed that both participants were complete strangers, and instructed them not to speak or otherwise interact with the each other prior to beginning the experiment. They were then fitted with motion capture suits and optical markers in accordance with OptiTrak's Baseline + 13 Skeleton. Each participant then played through the Oculus Touch Tutorial [38] to familiarize them with the Oculus Rift head-mounted display (HMD) and Touch controllers.

Participants were then told the rules of the tasks, positioned in the motion capture area, and fitted with the HMD and controllers (for VR conditions). Before the first task of each condition, participants played a short bluffing game (1 minute) to familiarize themselves with interacting in the condition. At

the conclusion of the second task, participants were given a survey to obtain their impressions of the task outcomes and the communication medium. This process was repeated for each of the remaining conditions.

At the conclusion of the third condition, participants were given one additional survey to gather information about their most and least favorite communication medium. Then, the researcher performed an experimental debrief with both participants, and encouraged the participants to discuss their survey answers and their general impressions of all three conditions.

Condition Implementation and Data Collection

For all conditions, subjects were recorded with three GoPro cameras. Audio was recorded either with lapel microphones or through the HMD microphone. For the VR conditions, the POV (Point of View video) of each participant was recorded during the interaction. In addition, the various transformations of each object within the scene were recorded at 20 frames per second, allowing us to create videos post-hoc of the interaction with color-coded avatars, including making avatars visible that had been hidden during the No Embodiment condition. These videos were used for later analysis.

Face-To-Face (F2F)

In the face-to-face condition, participants performed the tasks facing each other from across a table (Fig. 1-A and Fig. 1-B). Furniture was represented by foam boxes with full-color orthographic images of the furniture model on each side.

Virtual Reality with Full Embodiment (embodVR)

In the Full Embodiment VR condition, participants appeared inside of a grey room, on opposite sides of a table containing an apartment floor plan (Fig. 1-C); in actuality, participants were located on opposite sides of the motion capture space, facing opposite directions (Fig. 1-D). Table, furniture, and floor plan dimensions matched those of the face-to-face condition. Participants and their VR devices were tracked with a 24 camera (Prime 17W) OptiTrack motion capture system. Their positioning and pose was used to drive avatars and cameras within a customized Unity scene. The lag of the entire system was under 50 milliseconds; none of the participants mentioned noticeable lag during their exit interviews.

The HMDs employed were Oculus Rifts: immersive, head-mounted displays with 2160x1200 resolution, 90 Hz refresh rate, and 110° FOV [6]. See the supplementary video for examples of participants wearing the devices. The Rifts employed integrated microphones and in-ear earbuds to block out ambient noise and transmit mono audio between participants (via a local VoIP server). Participants also used hand-held Oculus Touch controllers to pick up furniture and make various hand shapes (fists, thumbs-up, pointing with index finger).

Virtual Reality without Embodiment (no_embodVR)

The No Embodiment VR condition was almost identical to the Full Embodiment VR condition. In the first task, however, neither avatar was visible to participants. In the second task, participants could view their own hands to assist in picking up furniture, but could not see any part of their partner's avatar nor the rest of their body. The workspace was fully visible.

Gesture Type	Description
Reference Object or Location	Deictic (or pointing) gesture to an object or location.
Reference Person	Deictic gesture at self or interlocutor.
Spatial or Distance	Gestures conveying more complex spatial or distance information, such as a path through the apartment.
Backchannel	Acknowledgments of interlocutor, including head nods and manual gestures.
Representation	Metaphoric and iconic hand movements, illustrative of an idea (but not fitting in "Spatial or Distance").
Emotional or Social	Gestures conveying strong emotions or other social information.
Beat	Small movements of the hand in rhythm with the spoken prosody.
Self-adaptor	Self-manipulations not designed to communicate, such as nose scratches.

Table 1. Annotators would apply one or more of these tags to each observed gesture.

MEASURES, RESULTS AND DISCUSSION

To measure the effects of virtual reality and embodied avatars on participant interaction, we employed several different types of measurements. For readability, we will group each measure description with related results and discussion in the sections below. Results are summarized in Table 3 and Figure 2.

ANNOTATED PARTICIPANT BEHAVIOR

Measure

Following the trials, a remote team annotated verbal and non-verbal behaviors exhibited by each dyad during the floor-plan negotiation task. This provided objective data on communication patterns to complement the subjective measures.

Annotators annotated each gesture performed by each participant, labelling its type. Following McNeill's [34] assertion that gesture should not be viewed categorically, but as having levels of different dimensions, annotators were allowed to apply more than one tag to a gesture. The gesture types are based on McNeill's [33] proposal of *deictic*, *beat*, *iconic* and *metaphoric*, but some dimensions were either collapsed or subdivided to focus the analysis on the most relevant behavior for the tasks conducted here. Gesture types are shown in Table 1.

Gesture may be redundant with or provide information not available in the verbal channel. As a simple example, consider a person pointing at room A on the floor plan. If they say "I want room A," the gesture is (largely) redundant. If they say "I want this room," the utterance cannot be understood without the gesture. Annotators were instructed to add a *Novel Content* tag to any gesture that contained information not available through the verbal channel.

Participants' dialog was annotated at two levels of granularity, as summarized in Table 2. *Utterances* are individual sentences or sentence-like units of dialog. *Conversational turns* denote a period when one person holds the floor before it passes to the other and may contain one or more utterances.

Speech Data	Description
Utterance	A section of speech. A sentence or comparable.
Pragmatic	Task related suggestions and discussion.
Social or Emotional	Strongly social or emotional utterances, such as “I’m very excited."
Non-task Discussion	Discussion not related to the task.
Backchannel	Verbal acknowledgements that indicate listening, such as “uh huh".
Complete Reference	Fully qualified references that can be completely understood from the utterance, like “I’d like room A".
Reference Pronoun	The use of terms like “this" or “that" to refer to things, such as “I’d like this room."
Conversational Turn	The duration for which one person holds the floor before the other takes over. Labeled with how the person gets the turn.
Interruption	The person takes the floor by interrupting the other.
No Marker	No clear indication of how the floor was obtained.
Verbal Hand Over	The interlocutor verbally passed the floor to the speaker.
Nonverbal Hand Over	The interlocutor nonverbally passed the floor to the speaker.

Table 2. Speech is tagged in the two levels specified, with individual tags listed below each level.

In the Face-To-Face condition, these annotations were made based on audio and video feeds from three camera angles. For the virtual reality conditions, annotations were made based on video feeds, audio and color-coded audio waveforms, POV footage for each participant, and multiple scene reconstructions in which avatars were always visible and color-coded.

To minimize the effects of individual annotators, all three of a dyad’s task conditions were annotated by the same annotators. To ensure high-quality annotation data, all tasks were annotated independently by two different annotators. Mismatches between the two annotators were resolved by a third annotator and quality checks were performed by the research team.

Results

A similar statistical approach was used for all data reported in this section. Repeated measures ANOVAs were run to determine if each dependent value varied significantly across the three conditions of F2F, embodVR and no_embodVR. Mauchly’s test for sphericity was run on all data and correction by Greenhouse-Geiser and Huynh-Feldt were applied as needed (both of these always succeeded). Type II error was corrected for using False Discovery Rate correction. When significant variation was found in the ANOVA, Bonferroni-corrected pairwise t-tests were run to determine which factors varied. Significance was evaluated at the $p < 0.05$ level.

Analysis of utterances focused on the distribution of utterance types. Their frequency is shown in Figure 2e. Condition had a significant effect on the occurrence of pragmatic utterances, with pragmatic utterances occurring significantly less frequently in the embodVR condition than in F2F. Condition also had a significant effect on the use of referential pronouns, with significantly fewer referential pronoun uses in the no_embodyVR condition than in F2F and embodVR.

Analysis of conversational turns focused on the frequency of conversational turns and the manner the turn was begun. Condition had a significant effect on the frequency of turns, with significantly fewer turns occurring in the no_embod condition than in the embodVR condition. The data shows a tendency for the same relationship between F2F and no_embodVR conditions ($p = 0.097$). The relative frequency of the manner by which a conversational turn began is shown in Figure 2f. Condition had a significant effect on the frequency of interruptions, with interruptions occurring more frequently in the F2F condition than in either embodVR or no_embodVR.

Analysis of nonverbal behavior focused on the frequency of gesturing, types of gestures employed and novelty of information carried by the gestures. The analysis showed that both F2F and embodVR had significantly higher gesturing rates than no_embodVR, but there was no significant difference between them (Figure 2a). In many dyads, one partner gestures more than the other. In no_embodVR, the less frequent gesturer made about 30 percent of the gestures, compared to 40 percent for F2F and embodVR. This disparity was significant between F2F and no_embodVR, and embodVR and no_embodVR, but not between F2F and embodVR. See Figure 2g.

The frequency of gesture types is summarized in Fig. 2d. Since a single gesture may display features of more than one type, the total of frequency counts may exceed 100% (in practice, it was ~120%). References to objects or locations were the most frequent, followed by representational gestures and gestures displaying complex spatial or distance information. The categories “Reference Person" and “Emotional or Social" never occurred more than 5% of the time and were dropped from the analysis. All remaining categories, except “Backchannel," showed significant differences. In every case except for “Self-adaptors", there was no significant difference between F2F and embodVR, but both of these differed significantly from no_embodVR. A higher proportion of gestures in no_embodVR were representational or beats and a lower proportion were object/location references and spatial/distance gestures. Self-adaptors were a higher proportion of gestures in F2F and no_embodVR, compared with embodVR, with no significant difference between these two categories.

Novel content was analyzed both in terms of the proportion of gestures that were so tagged and the number of novel content gestures per minute (Figure 2c). In both cases, there were no statistical differences between F2F and embodVR, but gesture with novel content was significantly lower in no_embodVR.

Discussion

Overall, the verbal behavior is more consistent across conditions and the nonverbal behavior shows greater variation, reflecting in part the visual nature of the nonverbal channel. Floor management is largely accomplished using nonverbal cues, such as gaze and posture [27], so it is more difficult over audio-only channels. The lower turn frequency in no_embodVR likely reflects the difficulty of efficiently obtaining and relinquishing the floor in this condition.

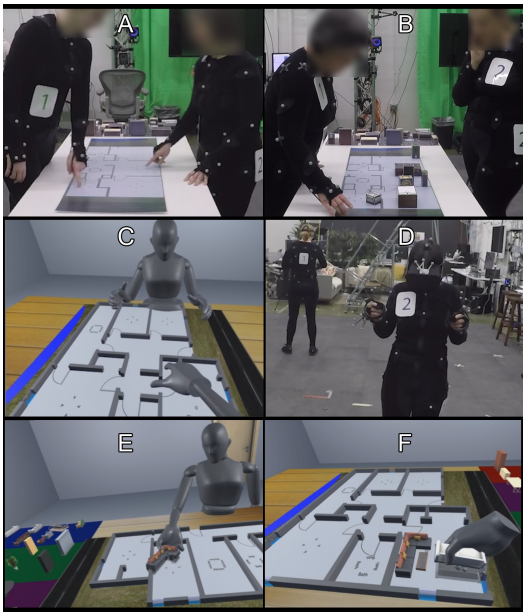

Figure 1. Dyads performed the first (A) and second (B) tasks in the face-to-face conditions. In virtual reality conditions, avatars appeared across the table from each other (C), but were actually positioned on opposite sides of the motion capture stage (D). In the embodVR condition, participants were able to see both avatars (E). In the no_embodVR condition, participants were unable to see their partner and could only see their hands in the second task, to assist with furniture manipulation (F).

The largely consistent utterance behavior suggests people are not making major changes in their conversational style, either to accommodate for the lack of nonverbal information in no_embodVR or due to any perceived differences between F2F and embodVR.

Reference pronouns such as 'this' or 'that' often require a gesture or visual indication to clarify their meaning. It is therefore reasonable that they occur significantly less frequently in the no_embodVR condition where there is no way to accompany the utterance with a visible gesture. Interestingly, participants used such pronouns at the same rate in the F2F and embodVR conditions, suggesting participants felt comfortably able to clarify their pronoun usage though avatar gesturing.

The dominance of pragmatic utterances across all three conditions indicate that participants remained focused on their task. However, there was a significant decrease in pragmatic utterances between the F2F and embodVR conditions. This may partially be explained by slightly more backchanneling and non-task discussions in the embodVR condition, though neither of these were not significant ($M_{F2F} = 29.9$, $M_{embodVR} = 32.5$, $p_{adj} = 0.29$).

The higher rate of interruptions in the F2F condition may reflect that participants have more visual information on their interlocutor (gaze, facial expressions) and hence have a greater sense of when it is possible to take the floor.

It was anticipated that people would gesture less in no_embodVR, when they can neither see themselves nor their partner. It is perhaps surprising that they still averaged 9.1 gestures per minute. More interesting is that there is no significant difference between the gesture rate in F2F and embodVR. This suggests that people gesture at more or less normal rates in embodied VR, even given the limitations of holding Touch controllers. The greater disparity in gesture rate within dyads for no_embodVR suggests that there may be entrainment behavior that occurs when people have visual access to their partner. Such entrainment is a feature of rapport, and visually displaying it during an interaction may be a way to increase the felt rapport [45].

It is reasonable that people make a significantly lower proportion of gestures that are tied to the visual environment ("Reference Object" and "Spatial") in no_embodVR. Similarly, they make a higher proportion of gestures without environmental references ("Representation" and "Beats"). Nonetheless, people still made referential gestures almost four times a minute in no_embodVR. These were often redundant, such as accompanying the utterance "I'd like room A." with a gesture pointing to the room, but when they weren't they could generate substantial confusion. It is again important to note that, aside from self-adaptors, there were no significant differences between F2F and no_embodVR, offering further evidence that normal communication patterns transfer over to embodied VR.

We did not develop a measure for gesture complexity, but after viewing the corpus, it appears that the gestures people make in no_embodVR are less complex and often smaller. The spatial gestures were generally the most complex in the corpus and often involved illustrating traffic flow in the apartment, how noise might travel or the relative location of rooms. While spatial gestures still occurred, none of these particularly complex forms were observed in the no_embodVR. It also does not appear that this level of detail was transferred to the verbal channel. Rather, some details of the arguments were left out.

To better understand the variation in self-adaptor behavior, we conducted a follow-up analysis looking at self-adaptor rate. They occurred 1.4 times per minute for F2F, 0.54 for embodVR and 0.68 for no_embodVR. The F2F rate was significantly higher than the other two. Self-adaptors are associated with anxiety and the personality trait of neuroticism [46, 3, 14, 11], so one possible explanation is that people are more anxious standing across the table from a flesh-and-blood person than they are in VR. It is also possible that they are more engaged in VR, so manipulate less, or the Touch Controllers make it more difficult to perform self-manipulations, although these occurred and there was some additional adjusting of the HMD.

With regards to novel content, once again F2F and embodVR show comparable performance. It is reasonable to expect novel content to be lower in no_embodVR as people make a conscious decision not to encode information on a channel that cannot be seen. The fact that people are continuing to encode novel content in gesture when not seen means that part of their message is lost, which can lead to misunderstandings.

It is interesting to note that despite there being no physical limits on people's movements in the embodVR, they respected standard rules of proxemics, but would occasionally move into each others space in no_embodVR where there was no indication of a body.

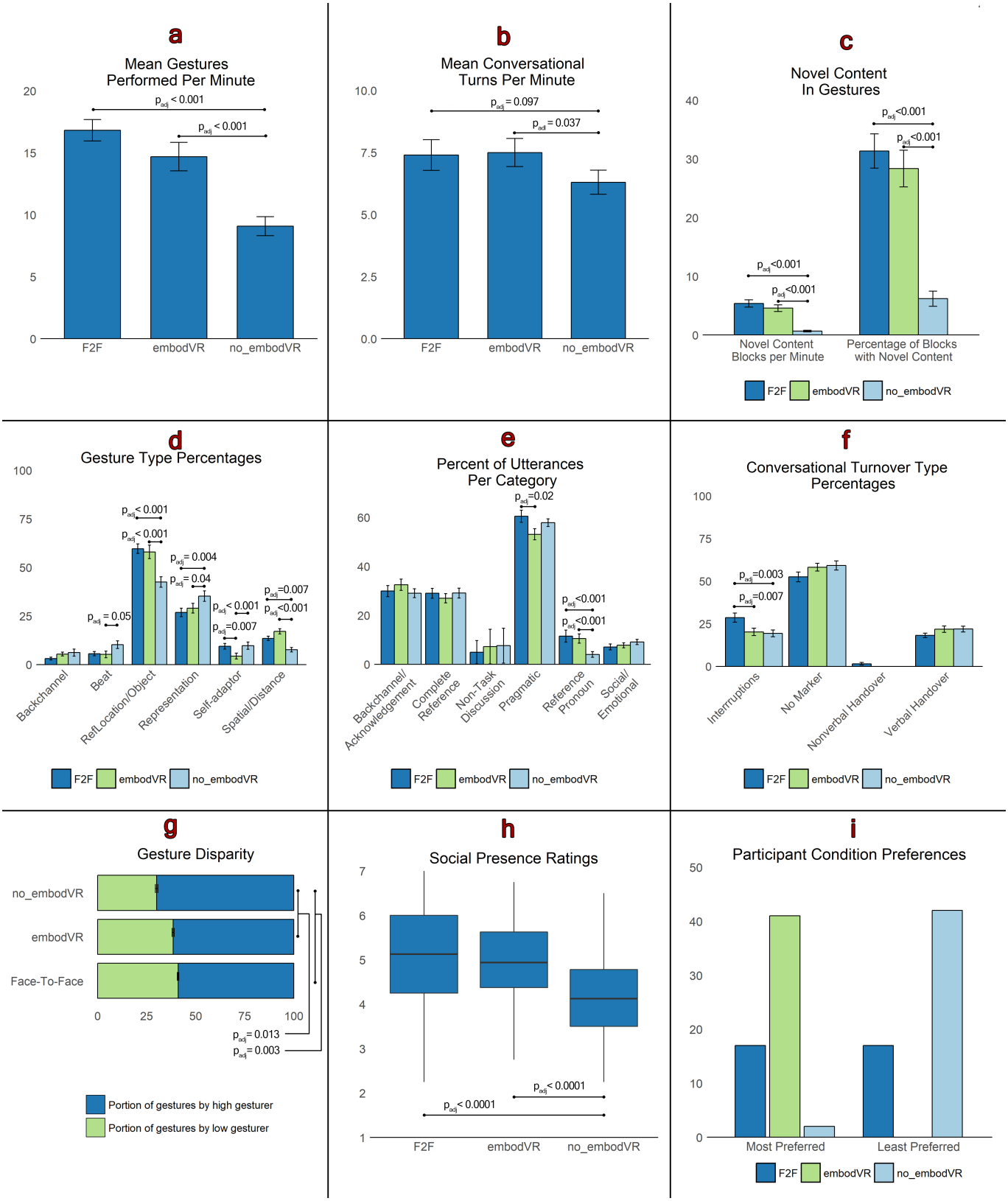

Figure 2. Sub figure a shows the average number of gestures performed per minute for each condition. Subfigure b shows the percentage of gestures that fall into each annotated category (note that, because some gestures fit multiple categories, totals for each condition can add up to over 100%). Subfigure c shows the rate and percentage of gestures which introduced novel content into the discussions (for example, point at a location while referring to it by a referential pronoun). Subfigure d shows the mean number of conversational turns taken per minute. Subfigure e shows the percent of utterances that fall in each annotated category (note that, because some gestures fit multiple categories, totals for each condition can add up to over 100%). Subfigure f shows the frequencies of the manners by which conversational turns were started. Subfigure g shows the ratio of gestures performed by the more frequent gesturer and less frequent gesturer in each dyad. Subfigure h shows the mean social presence scores, with standard errors of the mean, as measured by the semantic difference questionaire. Subfigure i shows the most and least favorite conditions, as reported by participants at the end of the experiment. All error bars show standard error of the mean.

SEMANTIC DIFFERENCE MEASURE OF SOCIAL PRESENCE

Measure

Social presence, or the sense of interacting with another, is a key factor in communication and a long term goal for virtual reality systems. To measure the degree of social presence afforded by each condition, participants completed a semantic difference survey immediately after completing the second task in each condition. Similar to previous works, our survey consisted of eight bipolar word pairs (e.g. "cold-warm", "impersonal-personal", "colorless-colorful") selected from [44]. Using a seven-point Likert scale, participants rated the degree to which they felt each adjective in the pair described the communication medium. This is a common technique and previous studies have found that communication mediums with higher degrees of social presence are often rated as warmer, more personal, and more colorful [23, 42, 36]

Results

A reliability analysis was conducted on the results of the semantic difference surveys by calculating Chronbach's alpha, which yielded a good internal consistency of 0.82. An average social presence score was then calculated from the factor responses for each participant and each condition. The mean and standard error are shown in Fig. 2h. Results of a repeated measures ANOVA, followed by pairwise comparisons using paired t-tests, indicate that both F2F and embodVR showed significantly higher perceptions of social presence than no_embodVR (medium effect size, Cohen's d of 0.62 and 0.65 respectively). There was no significant difference between F2F and embodVR (negligible effect size).

Discussion

Semantic Difference Measure of Social Presence

While it is not surprising that no_embodVR showed the lowest social presence, we expected F2F to still show greater social presence than embodVR, especially given that the current avatar is relatively primitive, lacking facial expressions, a full set of hand movement, muscle deformations, etc. Despite this, both the results and comments in the surveys and exit interview seem to indicate that people felt a high level of social presence with their interlocutor when the avatar was present.

NETWORKED MINDS MEASURE OF SOCIAL PRESENCE

Measure

Following the semantic difference survey, participants were asked to reply to an additional 36 prompts on a 7-point Likert scale (Please see supplemental material.). These questions were based on the Networked Minds survey [9, 8, 7], an alternative measure of social presence, as well as including additional items deemed relevant for this study.

Results

In the data from the long survey, 6 cells (of 6,480) were blank because subjects forgot to circle answers. These blanks were replaced with the average score of all participants in that condition. In addition, one subject missed 24 questions in the no_embodVR condition. Data for that subject and condition was excluded from the analysis.

A factor analysis was performed on the full set of questions, described in detail in the supplemental material. It yielded six factors, four of which were maintained: Clarity of Communication, Social Awareness, Conversation Management, and Disconnection to Partner. Chronbach's alpha for these four factors produced alpha's of 0.92, 0.86, 0.81, and 0.76 respectively.

ANOVAs showed that all four factors were significantly affected by condition at the $p < 0.05$ level, as summarized in Table 3 and detailed in the supplemental document. Post-hoc analysis showed that there was no significant difference between embodVR and F2F on Clarity of Communication, Conversation Management and Disconnection to Partner. For Conversation Management, embodVR performed significantly better than no_embodVR. F2F and embodVR performed significantly better than no_embodVR on the other two factors, showing the same pattern as with the semantic difference measure. Effect sizes were medium in each case, except between F2F and no_embodVR, where it was small. For Social Awareness, a three-level order appears where F2F performs better than embodVR which performs better than no_embodVR, with means of 6.16, 5.81 and 4.31 respectively. There was a medium effect size between F2F and embodVR (Cohen's d = 0.51). There was a large effect size between both F2F and embodVR when compared to no_embodVR (Cohen's d of 1.3 and 1.1 respectively).

Discussion

Three of the factors, including connection with partner, showed the same pattern as semantic differencing, with no significant difference between F2F and embodVR, but a degradation for no_embodVR, offering further evidence that people experienced similar social presence in F2F and embodVR. For Social Awareness, the degradation had a large effect size when comparing either F2F or embodVR with no_embodVR, but there was also a medium size difference between F2F and embodVR. Looking at the individual components, the factors "I could tell what my partner was paying attention to." and "I am confident I understood the emotions expressed by my partner." have the largest impact on this difference between F2F and embodVR. Gaze and emotions are highly dependent on eye movement and facial expressions, both missing from embodVR, so this may explain its lower performance to F2F on this factor.

PARTICIPANT PREFERENCES AND EXIT INTERVIEWS

Measure

At the conclusion of the final condition, participants were asked to list their favorite and least favorite communication medium in the experiment, along with reasons for their answers. Following the debrief, informal exit interviews were conducted with the last 52 participants. These began with a prompt such as "Do either of you have any questions for me or any impressions you'd like to share?".

Results

Results for participants most and least preferred interface are shown in Fig. 2i. EmbodVR was most preferred by 39 participants and least preferred by 0. F2F was most preferred

by 15 and least preferred by 17. Four people selected both embodVR and F2F as their favorite. No_embodVR was most preferred by 2 and least preferred by 42.

Discussion: Participant Preferences

To gain a deeper understanding of the preference results, we categorized the written reasons given for people's most and least preferred interfaces. One obvious explanation for the preference of embodVR is novelty, and there is some evidence for this. Of those who preferred embodied VR, five mentioned something that related to novelty in their explanation. For those that least-preferred face-to-face, the lack of novelty came up for seven participants, many of whom thought it was "boring". Novelty does not seem to be a complete explanation, both because no_embodVR is also novel and the many other justifications offered.

Of those who preferred embodied VR, eight mentioned they liked seeing their interlocutor, seven thought the interface was fun and/or exciting, six explicitly mentioned the importance of body language and being able to show things in the environment, with comments like "I got tone and body language." Five people thought the interface was more personal and social. Three of the people who preferred embodied VR mentioned a sense of presence, with comments "I enjoyed seeing my partners body language/movement. I felt like she was in room with me even though she wasn't", "[It] allowed me to interact on a personal level with a stranger while still being intimate and fully immersed", and even "seeing and hearing a real person is amazing".

Of those that least preferred no_embodVR, 12 commented that it felt impersonal, sterile or they had less connection with their partner. For example, saying "it felt the most distant and least like there was another person across from me."

For some, the abstraction of a grey, faceless avatar used in the embodVR condition increased their comfort with the interaction. Four people mentioned this who listed the condition as their favorite, with comments "being in VR makes it somehow less intimidating arguing over room space when you don't know your partner well", "[I] felt less shy and self-conscious than I would have otherwise because of this interface", "[embodVR was best] because it was a fun and safe environment to navigate things I wanted. In-person it felt awkward [F2F], last round [no_embodVR] it just felt like I could do it by myself. I didn't have a connection to my partner." Another commented, "the [face-to-face was my least favorite] for sure. I do not like conflict and in the [two VR conditions] it was easier to voice my opinion."

We anticipated that the lack of facial expressions would be an issue. There is some evidence that it was in the surveys, but less than anticipated. Two people who mentioned embodied VR as their preferred interface said that they felt the lack of facial expressions, e.g. "But lack of facial expression made hard to know his actual feeling". Two people that preferred face-to-face mentioned facial expressions in their explanation. Three people that preferred face-to-face also mentioned that they it was easier to see expressions and understand their partner, which may include facial expressions.

Discussion: Exit Interviews

The exit interviews are particular useful for understanding the more social aspects of people's experience with the system. Most people felt that having an avatar improved communication over the no_embodVR condition (28/52 mentioned this). For example, P10 commented "I think by seeing the avatar ... this changes the entire outcome of the conversation....I think based on body language ... it makes it easier to have a communication and to find an agreement." P59 noted how the body focused her attention: "Actually being able to see another body across the table, your focus ... you're automatically drawn into that situation and you're way more focused. And it is amazing how much being able to gesticulate, having that ability to gesticulate, how that is so much part of the communication process....". The richness of embodied communication and how it impacted decision making was highlighted by P30 "I was more influenced by his body language in [embodVR], because I could see if he really liked it, or he was just making an effort to make something different."

In looking at the mechanisms that helped in communication, some people mentioned the role of gestures and pointing. Three people pointed out that the embodied avatar allowed them to anticipate what the other person was thinking earlier and prepare a response, for example "It was nice to ...actually see the person and then seeing them actually begin reaching to an object, so you can .. you're already thinking of your reaction to it, what you're going to say,... so you're preemptively thinking of how am I going to speak to him about it, like what am I going to say, how am I going to disagree, like 'oh. I knew that piece, I didn't like that piece', alright, I'm already thinking about how I'm going to disagree with him."[P36]

People noticed the lack of facial expressions in the avatar, with ten participants suggesting that adding facial expressions would be helpful and three noting the need for eye contact. Five participants commented that it was easier to read expressions in the face-to-face condition.

Participant comments suggested that they feel more alone and cold in the no_embodVR condition and a much greater sense of social presence with the avatar (44 of 52 participants commented on this in some form). Eleven participants commented that it felt like the other avatar was standing directly in front of them, even though they knew the person was physically standing on the other side of the room and faced away from them. Indicative comments include: P60 saying "I felt like you could see me. That was the weird thing....That changes your behavior." and P59 replying "I really felt like I was talking to that grey thing." P21 commented, "I felt like the avatar was really interesting, because even though I couldn't see her facial expression, I could see her body movement and I felt her. I felt her presence there."

A surprising finding is some evidence that people changed the competitiveness of their interaction depending on the interface used. They felt they were being more considerate and empathetic when a body was present than in no_embodVR, where they were willing to be more competitive, aggressive and less likely to compromise. Nine participants made comments related to the no_embodVR interface depersonalizing

Category	Factor	ANOVA Result	Post-hoc
Semantic Difference	Avg. Semantic Difference Score	$F_{2,116} = 18.07, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
Detailed Survey	Clarity of Communication	$F_{2,116} = 12.13, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
	Social Awareness	$F_{2,116} = 91.30, p_{adj} < 0.001$	$F2F > embodVR > no\_embodVR$
	Conversation Management	$F_{2,116} = 4.89, p_{adj} = 0.012$	$embodVR < no\_embodVR$ (lower is better)
	Disconnection to Partner	$F_{2,116} = 13.39, p_{adj} < 0.001$	$F2F, embodVR < no\_embodVR$ (lower is better)
Utterance Type	Pragmatic	$F_{2,56} = 4.36, p_{adj} = 0.043$	$F2F > embodVR$
	Social/Emotional	$F_{2,56} = 1.62, p_{adj} = 0.260$	No significance
	Non-Task Discussion	$F_{2,56} = 0.89, p_{adj} = 0.500$	No significance
	Backchannel	$F_{2,56} = 1.75, p_{adj} = 0.260$	No significance
	Complete Reference	$F_{2,56} = 0.61, p_{adj} = 0.547$	No significance
	Reference Pronoun	$F_{2,56} = 13.04, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
Turn Frequency	Conversational Turn Frequency	$F_{2,56} = 3.94, p = 0.025$	$embodVR > no\_embodVR$
Turn Type	Verbal Handover	$F_{2,56} = 1.77, p_{adj} = 0.179$	No significance
	Nonverbal Handover	$F_{2,56} = 3.65, p_{adj} = 0.064$	No significance
	Interruptions	$F_{2,56} = 6.9, p_{adj} = 0.001$	$F2F > embodVR, no\_embodVR$
	No Marker	$F_{2,56} = 2.36, p_{adj} = 0.138$	No significance
Gesture Behavior	Gesture Frequency	$F_{2,58} = 34.75, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
	Gesture Disparity	$F_{2,58} = 9.83, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
Gesture Type	Reference Object or Location	$F_{2,58} = 18.91, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
	Spatial or Distance	$F_{2,58} = 14.01, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
	Backchannel	$F_{2,58} = 2.06, p_{adj} = 0.14$	No significance
	Representation	$F_{2,58} = 6.60, p_{adj} = 0.004$	$F2F, embodVR < no\_embodVR$
	Beat	$F_{2,58} = 5.46, p_{adj} = 0.008$	$no\_embodVR > embodVR$
	Self-adaptor	$F_{2,58} = 7.78, p_{adj} = 0.002$	$F2F, no\_embodVR > embodVR$
Novel Gesture Content	Percent of Blocks with Novel Content	$F_{2,58} = 53.94, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$
	Novel Content Blocks Per Minute	$F_{2,58} = 46.27, p_{adj} < 0.001$	$F2F, embodVR > no\_embodVR$

Table 3. Results from ANOVAs and significant post-hoc t-tests for all computed measures. Verbal and nonverbal measures are calculated on the annotation data from the floor plan task. Significance values for post-hoc results are reported in Figure 2

the interaction. Four commented that they were more aggressive, more direct, more willing to argue or less compromising in no_embodVR. Two of these found the task easier in VR because it was depersonalized. Two other subjects thought they were more objective in the avatar condition, especially without a body. Two other participants felt less present in no_embodVR so wanted to control the situation more. A different dyad, outside of the nine, expressed that it was easier to "bicker" when not in face-to-face. One other subject felt that body language made it easier to reach agreement. An additional person commented that turn taking was more formal in no_embodVR.

Some participants felt more comfortable in VR than in F2F. For example, P55 said "I just felt like it was easier for me to be more relaxed and more myself [in VR] because ... I don't know ... it just gave me a safe place to do it ... like, in person, I wouldn't want to upset you, but when it became virtual reality it was a little different ... like ... I don't know, I just felt like I relaxed more." This could be related to some participants feeling less revealed in VR, even if overall social presence was similar.

CONCLUSIONS

Embodied virtual reality and face-to-face interaction showed remarkably similar verbal and nonverbal communicative behavior, with the anticipated drop off for VR without bodies. Having a tracked body in the virtual world seems to help people feel that they are really interacting with another person: all but one subjective measure showed no significant difference for social presence between F2F and embodVR, with lower social awareness possibly reflecting the lack of facial information. There was a clear preference for including a body in the experience as people felt "alone" in no_embodVR and ratings dropped. Removing the body decreased referential pronoun usage and lowered the frequency with which participants took conversational turns.

There are, of course, limitations to the work. The first is that this study examined a particular context in which users have a shared visual work space. The activities included a negotiation task and a design task. Behavior may vary for different environments and different activities. A second limitation is that while we measure conversational behavior and subjective experience, we don't measure the effectiveness of the conversation. Both of these issues point to interesting follow-up work. For example, it would be interesting to examine social conversation to see whether facial motion plays a more dominant role here. Facial animation was excluded from this study both due to technical limitations and in order to focus on the impact of body movement. The study also used relatively low-fidelity models. It would be interesting to see if behavior and experience changes with photo-realistic models that include facial animation.

Acknowledgements

We gratefully thank the team at Oculus Research. In particular, this work would not have been possible without Ronald Mallet, Alexandra Wayne, Matt Vitelli, Ammar Rizvi, Joycee Kavatur and the annotation team.

REFERENCES

1. U.S. Department of Energy Advanced Research Projects Agency Energy (ARPA-E). 2017. FACSIMILE APPEARANCE TO CREATE ENERGY SAVINGS (FACES). (2017).
2. Leila Alem and Jane Li. 2011. A study of gestures in a video-mediated collaborative assembly task. *Advances in Human-Computer Interaction* 2011 (2011), 1.
3. M. Argyle. 1988. *Bodily communication*. Taylor & Francis.
4. Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).
5. Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and virtual environments* 10, 6 (2001), 583–598.
6. Atman Binstock. 2015. Powering the Rift. (2015). `https://www.oculus.com/blog/powering-the-rift/`
7. Frank Biocca, Judee Burgoon, Chad Harms, and Matt Stoner. 2001. Criteria and scope conditions for a theory and measure of social presence. *Presence: Teleoperators and virtual environments* (2001).
8. Frank Biocca, Chad Harms, and Judee K Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and virtual environments* 12, 5 (2003), 456–480.
9. Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence, Philadelphia, PA*. 1–9.
10. Sara A Bly. 1988. A use of drawing surfaces in different collaborative settings. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*. ACM, 250–256.
11. A. Campbell and J. Rushton. 1978. Bodily communication and personality. *The British Journal of Social and Clinical Psychology* 17, 1 (1978), 31–36.
12. Herbert H Clark, Susan E Brennan, and others. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.
13. Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language* 50, 1 (2004), 62–81.
14. P. Ekman and W. V. Friesen. 1972. Hand movements. *Journal of Communication* 22 (1972), 353–374.
15. Susan R Fussell, Robert E Kraut, and Jane Siegel. 2000. Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 21–30.
16. Susan R Fussell and Leslie D Setlock. 2014. Computer-mediated communication. *Handbook of Language and Social Psychology. Oxford University Press, Oxford, UK* (2014), 471–490.
17. Susan R Fussell, Leslie D Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam DI Kramer. 2004. Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction* 19, 3 (2004), 273–309.
18. Darren Gergle, Robert E Kraut, and Susan R Fussell. 2004a. Action as language in a shared visual space. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. ACM, 487–496.
19. Darren Gergle, Robert E Kraut, and Susan R Fussell. 2004b. Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of language and social psychology* 23, 4 (2004), 491–517.
20. Darren Gergle, Robert E Kraut, and Susan R Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction* 28, 1 (2013), 1–39.
21. Darren Gergle, Carolyn P Rosé, and Robert E Kraut. 2007. Modeling the impact of shared visual information on collaborative reference. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1543–1552.
22. Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 413–422.
23. Jörg Hauber, Holger Regenbrecht, Aimee Hills, Andrew Cockburn, and Mark Billinghurst. 2005. Social presence in two-and three-dimensional videoconferencing. (2005).
24. Aimèe Hills, Jörg Hauber, and Holger Regenbrecht. 2005. Videos in space: a study on presence in video mediating communication systems. In *Proceedings of the 2005 international conference on Augmented tele-existence*. ACM, 247–248.
25. Ellen A Isaacs and John C Tang. 1994. What video can and cannot do for collaboration: a case study. *Multimedia systems* 2, 2 (1994), 63–73.
26. Steven Johnson, Madeleine Gibson, and Bilge Mutlu. 2015. Handheld or handsfree?: Remote collaboration via lightweight head-mounted displays and handheld devices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1825–1836.
27. Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63.

28. David Kirk and Danae Stanton Fraser. 2006. Comparing remote gesture technologies for supporting collaborative physical tasks. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 1191–1200.
29. Robert M Krauss, Connie M Garlock, Peter D Bricker, and Lee E McMahon. 1977. The role of audible and visible back-channel responses in interpersonal communication. *Journal of personality and social psychology* 35, 7 (1977), 523.
30. Robert E Kraut, Susan R Fussell, and Jane Siegel. 2003. Visual information as a conversational resource in collaborative physical tasks. *Human-computer interaction* 18, 1 (2003), 13–49.
31. Robert E Kraut, Steven H Lewis, and Lawrence W Swezey. 1982. Listener responsiveness and the coordination of conversation. *Journal of personality and social psychology* 43, 4 (1982), 718.
32. Robert E Kraut, Mark D Miller, and Jane Siegel. 1996. Collaboration in performance of physical tasks: Effects on outcomes and communication. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. ACM, 57–66.
33. David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.
34. D. McNeill. 2005. *Gesture and thought*. University of Chicago Press.
35. Ian E Morley and Geoffrey M Stephenson. 1969. INTERPERSONAL AND INTER-PARTY EXCHANGE: A LABORATORY SIMULATION OF AN INDUSTRIAL NEGOTIATION AT THE PLANT LEVEL. *British Journal of Psychology* 60, 4 (1969), 543–545.
36. Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 481–494.
37. Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. 1993. Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-computer interaction* 8, 4 (1993), 389–428.
38. Oculus. 2016. Oculus Touch Tutorial. Oculus Store. (2016).
39. Gary M Olson and Judith S Olson. 2000. Distance matters. *Human-computer interaction* 15, 2 (2000), 139–178.
40. Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1964. *The measurement of meaning*. University of Illinois Press.
41. Tomislav Pejsa, Michael Gleicher, and Bilge Mutlu. 2017. *Who, Me? How Virtual Agents Can Shape Conversational Footing in Virtual Reality*. Springer International Publishing, Cham, 347–359.
42. Jan Richter, Bruce H Thomas, Maki Sugimoto, and Masahiko Inami. 2007. Remote active tangible interactions. In *Proceedings of the 1st international conference on Tangible and embedded interaction*. ACM, 39–42.
43. Abigail J Sellen. 1995. Remote conversations: The effects of mediating talk with technology. *Human-computer interaction* 10, 4 (1995), 401–444.
44. John Short, Ederyn Williams, and Bruce Christie. 1976. The social psychology of telecommunications. (1976).
45. Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1, 4 (1990), 285–293.
46. P. Waxer. 1977. Nonverbal Cues for Anxiety: An Examination of Emotional Leakage. *Journal of Abnormal Psychology* 86, 3 (1977), 306–314.
47. Steve Whittaker. 2003. Theories and methods in mediated communication. *The handbook of discourse processes* (2003), 243–286.
48. Steve Whittaker, Erik Geelhoed, and Elizabeth Robinson. 1993. Shared workspaces: how do they work and when are they useful? *International Journal of Man-Machine Studies* 39, 5 (1993), 813–842.