Videoconference and Embodied VR: Communication Patterns Across Task and Medium

AHSAN ABDULLAH, University of California, Davis, USA
JAN KOLKMEIER, University of Twente, The Netherlands
VIVIAN LO, Facebook Reality Labs, USA
MICHAEL NEFF, Facebook Reality Labs and University of California, Davis, USA

Fig. 1. Frames from the video log showing participants interacting over video conference (left) and in embodied virtual reality (right) while discussing an apartment floor plan.

Videoconference has become the dominant technology for remote meetings. Embodied Virtual Reality is a potential alternative that employs motion tracking in order to place people in a shared virtual environment as avatars. This paper describes a 210 participant study focused on behavioral measures that compares multiparty interaction in videoconference and embodied VR across a range of task types: a factual intellective task, a subjective judgment task and two negotiation tasks, one with visual grounding. It uses state-of-the-art body, face and finger tracking to drive the avatars in VR and a carefully matched videoconferencing implementation. Significant behavioral differences are observed. These include increased activity in videoconference related to maintaining the social connection: more person directed gaze and increased verbal and nonverbal backchannel behavior. Videoconference also had reduced conversational overlap, increased self-adaptor gestures and reduced deictic gestures as compared with embodied VR. Potential explanations and implications are discussed.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: VR, telepresence, remote meetings, gaze, nonverbal behavior, multiparty interaction

Authors' addresses: Ahsan Abdullah, aabdullah@ucdavis.edu, University of California, Davis, Davis, CA, USA, 95616; Jan Kolkmeier, University of Twente, Twente, The Netherlands, j.kolkmeier@utwente.nl; Vivian Lo, Facebook Reality Labs, Sausalito, USA, vivian.lo@fb.com; Michael Neff, Facebook Reality Labs and University of California, Davis, USA.

2573-0142/2021/10-ART453 $15.00
https://doi.org/10.1145/3479597

ACM Reference Format:
Ahsan Abdullah, Jan Kolkmeier, Vivian Lo, and Michael Neff. 2021. Videoconference and Embodied VR: Communication Patterns Across Task and Medium. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 453 (October 2021), 29 pages. https://doi.org/10.1145/3479597

1 INTRODUCTION

Videoconference, the dominant medium for remote meetings, uses video cameras to provide remote participants with a 2D, screen-based visual connection. Embodied VR, an emerging alternative, uses motion tracking and VR headsets to place participants in a shared 3D environment. This immersive 3D experience is more similar to face-to-face interaction, although at a lower fidelity, and it is important to understand how it impacts the collaborative experience. This paper describes a study comparing people's *behavioral patterns* across the two media and over a set of four tasks spanning different elements of workplace meetings. It builds on a long standing interest in how the affordances of communication media support various tasks [14, 19, 61]. Understanding the behavior engendered by the different affordances has important ramifications for the design of remote collaboration systems. As but one timely example, recent work postulates the exhaustion people feel from videoconferencing, so called "Zoom Fatigue", may result in part from the behavioral pattern of people receiving too much gaze [1].

The study employs state-of-the-art, embodied VR technology that includes body tracking, face tracking and finger tracking to drive the movement of semi-realistic avatars (Figure 1, accompanying video), which provides a compelling interaction experience. Great care was taken to ensure that the videoconference (VC) and virtual reality (VR) conditions were as evenly matched as possible in the experiment, for example by employing a recommended videoconference framing that shows the upper body so that gestures read clearly, providing a shared mouse interface so that people could still point at shared artifacts in VC like they can in VR, and having all participants maintain a fixed, seated position in both conditions. Differences remain, however. The model based avatars have lower fidelity than video and do not fully reveal a person's identity (gender and ethnicity were matched). Conversely, the avatars allow people to be located in a shared 3D space, while videoconference remains screen based.

Participants worked in groups of three to complete a warm up and four experiment tasks that were designed to replicate different types of activities that might occur during meetings. An *intellective task* required them to come up with answers to questions where there was a correct answer. A *decision making task* required them to reach consensus when there was not a single correct answer. Two *mixed-motive tasks* required them to negotiate where each team member had different desires. The second mixed-motive task introduced a floor plan to visually ground the task and explore how this impacted nonverbal behavior. The experiment was run with Medium as a between subjects condition and Task as a within subjects condition. In other words, every participant was assigned to only one Medium, but completed all four Tasks. In total, 210 people participated in the study, 35 groups of three in each medium.

An analysis of performance and subjective measures (e.g. social presence) did not reveal notable differences between the two media. For instance, post task surveys derived from [10, 25, 45, 60, 61] showed high ratings without significant differences between media for the scales *Satisfaction with Medium* (mean 6.30 for VC and 6.41 for VR on seven-point Likert scales), *Co-presence* (mean of 6.48 for VC and 6.53 for VR) *Mutual Understanding* (mean of 6.12 in VC and 6.24 in VR) and *Clear Communication of Affect* (mean of 5.42 for VC and 5.37 for VR) This paper focuses on the marked differences in behavioral measures.

Problem Statement: This work seeks to understand the behavioral differences that arise from people's use of either embodied VR or videoconference as a medium for conducting work meetings

and if these behavioral differences are based on the nature of the work task. As the current default remote meeting option, VC provides an important comparison point for evaluating the behaviors induced by embodied VR. It is important to understand the potential and impacts of embodied VR ahead of potential widespread adoption, and these are partially contained in the behavioral patterns the media encourages.

Contributions: As far as we are aware, this paper reports on the first large scale comparative study of behavioral patterns of triad interaction across videoconference and embodied VR. The study included 210 participants with diverse demographics. To illuminate the role played by the different technologies, the basic meeting configuration is kept as similar as possible between the two media. Different types of tasks are contrasted to explore if behavior changes as a function of task. Evaluated behavioral measures include conversational turns, gaze patterns and nonverbal behavior. *The paper contributes clear evidence of marked behavioral differences across task and media. Media differences include that VC participants spent much more time looking at interlocutors, especially their faces, they provided more verbal and nonverbal backchannels, performed more self-adaptors, fewer deictic gestures, and in some cases, had longer conversational turns. A potential explanation for these behavior changes is that participants in videoconference exerted greater effort to maintain a social connection than participants in embodied VR.* This suggests a differential in exertion required to use the two media that will likely impact users and warrants further in depth study.

2 BACKGROUND

2.1 Theory

Media Affordances and Nonverbal Communication: There has been a long term interest in studying how the affordances of different media impact conversational interaction (e.g. [14]). Conversation is a collaborative process in which meaning is incrementally constructed together. It relies on both coordination and communication across verbal and nonverbal channels. It is made more efficient through *grounding*, a process through which interlocutors develop a shared understanding, and this coordination can be easier with a shared environment [64].

Conversational turn management indicates when it is another person's turn to speak and is largely done nonverbally. Head nods, gaze, and gesture all mediate turn taking [64]. People use more words and turns in audio telephony than face-to-face communication, and most notable for this study, turn taking in video conferencing tends to be more formal compared to face-to-face communications and is more similar to that seen in telephone calls [20, 43]. It is postulated that the increased verbal communication is a compensation for visual grounding that is less effective than in face-to-face settings [20].

Backchannels are feedback that listeners provide to speakers indicating that they are paying attention, have understood what is being said, etc. They are often nonverbal and include actions like head nods [9] and also phrases like "mmm" and "mhmm". Backchannels lead to smoother communication between the speaker and listener. O'Conaill et al. [43] found more auditory backchannels in face-to-face meetings than an early, high quality videoconference system, and by far the fewest in a low quality, single duplex videoconference system.

Gaze plays a rich set of functions, including expressing intimacy, exercising social control, regulating interaction and providing information [34]. It communicates a person's attention. In face-to-face communication, it allows people to tell who is staring at whom [19]. Gaze duration and looking at another's face are powerful cues [64].

Deictic gestures establish reference by pointing at objects and assist with grounding. Nonverbal deixis can increase the efficiency of communication [64]. Gestures can make representations of

objects (iconic gestures) or ideas (metaphoric) [40]. Gestures are also used to regulate turn taking. They can be used to indicate emphasis, tone and subtext [33].

Finally, nonverbal communication performs a range of social functions, including: impression formation, person perception, communication of emotions and interpersonal attitudes [9] Bente et al. [9] argue that "[t]he effects emerge from implicit dynamic qualities, which rarely pass the threshold of conscious registration."

Task: Several taxonomies of group work have been put forward (e.g. [23]) and we rely on that of McGrath in designing our tasks [39, 61]. Their circumplex model consists of two dimensions, one runs from Conceptual to Behavioral; the second relates to the degree and nature of interdependence in three levels: collaboration, coordination, conflict resolution [61]. Following [61], social context cues are relatively unimportant when there is a demonstrably correct answer. The communication medium is more likely to have an effect when tasks require coordination, expression and perception of emotion, and persuasion or reaching consensus [61]. Whittaker argues "[f]or social tasks, there are clearly differences between mediated and face-to-face interaction, but for many cognitive tasks (especially those that do not require access to a shared physical environment), outcomes may not be different." [64] Tasks involving interdependence or uncertainty require substantial amounts of interpersonal communication to be successful [35], which in turn places demands on the quality of the communication medium. Our tasks were designed to span this range (Sec. 3).

2.2 Related Work

There is a large related literature that is briefly sampled here. Influential for this work is the study of Strauss and McGrath [61] that focused on how medium (online chat system or face-to-face) interacts with task type (idea generation, an intellective task, and a judgment task). Results showed very similar quality output across the two interfaces, but face-to-face was more efficient. In the judgment task, people were less productive for the mediated interface and responded more negatively to the medium and task.

An early ethnographic study of videoconferencing [30] showed it had advantages over audio only in factors like showing understanding, expressing attitudes and nonverbal communication, but performed worse than face-to-face for peripheral cues, controlling the floor and pointing to objects, with the lack of correct eye contact seen as distancing. Dong and Fu [18] found videoconference more successful than audio or text for negotiation and attributed the difference to exchanging information in small pieces. Hauber et al. [27] found that spatial interfaces based on using multiple video screens to create a 3D environment positively influenced social presence and copresence measures in comparison to 2D, but the task measures favored the two dimensional interface. Other work shows a benefit of adding spatial video to an audio conference [29]. Nguyen et al. [41] compared videoconferencing systems that had a single camera for 2-3 remote participants with ones that had dedicated cameras and projectors directed for each participant. The non-directed video condition showed significantly less co-operative behavior than either directional video or face-to-face. Follow-up work showed greater empathy for an upper-body video framing than head-only [42], so that wider framing was adopted in this study. Other work has explored video projections for two person remote interaction [48]. Schroeder [52] suggests that avatars could provide the spatial component missing in video, but they suggest a concern that the representation of the person may not be authentic. Wong and Gutwin [65] found that pointing in collaborative virtual environments benefited from being able to observe the prepatory arm motion, a direct connection between the gesturer and referent, and awareness of others views.

Early avatar research used a mixed head mounted displays with participants at workstations, and found a positive relationship between presence and co-presence, with accord increasing with presence [56]. Dodds et al. [16] found that a gesturing avatar led to more words guessed correctly

in a game scenario and more gestures than a static avatar. Bente et al. [9] conducted an early, large scale study of avatar representations in which pairs of participants selected job applicants using one of six interfaces: text, voice, video conferencing, low-fidelity avatar (cartoon-style) and high-fidelity avatar (3D character). Text performed worse than all conditions on perceived intimateness, co-presence and emotionally based trust. Most dependent variables did not show a difference between video and avatar conditions. Notably, their avatars were displayed on 2D screens rather than the shared 3D environments used in this study. Steptoe et al. introduced one of the first avatar gaze tracking systems and provide preliminary evidence that it improves communication [59]. They later found that realistic eye movement increases participants ability to detect truth and deception [58].

A preliminary evaluation of the Holoportation AR avatar system suggests participants experienced spatial and social presence, and appreciated being able to control their point of view [46].

Pan and Steed showed that people asked for less advice from a key-frame animated, 2D projected avatar than a video or a robot expert, but would always prefer the more expert agent [47]. Smith and Neff [57] showed similar social presence and behavioral patterns for faceless, motion tracked avatars and face-to-face communication, but lower presence and a shift in communication patterns when avatars were not present in VR. Other work found no differences for avatar rendering style, but some differences for the amount of the body that was motion tracked [67] and a preference for full body avatar motion [28]. Jo et al. [31] found that video performed worse than avatars on a measure that included spatial and social presence questions.

This study employs model-based avatars. A pre-rigged character model is used for each avatar, much like would be used in videogames, its movements driven by live tracking. An alternative avatar technology employs depth cameras or other optical techniques to create a point-cloud model of the person in real time, also known as 3D video (e.g.. [4, 22, 46] and hybrid approaches [32]). 3D video has the potential advantage of better preserving the person's identity, but tends to suffer from visual artifacts such as tear out (holes in the mesh), pixelation and issues with occlusion, as well as requiring complicated capture setups. It is also difficult to place multiple avatars in a shared 3D environment with this technique without also rendering the head mounted displays on the avatars, which blocks the face and limits communication. This paper is not focused on the particular avatar technology and we will simply note that at this time, model-based approaches offer more consistent visual quality and easier immersion in 3D.

The authors of impressive recent work suggest that it is the first to feature avatars with live tracking of the body, gaze and the lower portion of the face [50] and develop methods to augment avatar behavior beyond participants' actual motions. Our work tracks the same features, and also tracks hands and the full face, but our focus is on studying interaction patterns relative to a videoconference baseline, so we do not intentionally modify participant behavior. Other work has also explored the potential of adaptation, focusing on facial expressions in VR [26]. These studies point to the additional potential VR offers for modified or augmented social interaction.

3 METHOD

3.1 Experiment Design

Participants attended a single session, during which they interacted with two other participants in technologically mediated social interaction. The experiment design is mixed, with a between groups factor *mediation interface* at two levels: embodied VR and VC (see sec. 3.5), and a within group factor of *task type* at four levels (see sec. 3.2). The order of *task* was randomized. In short, groups of three completed all tasks with one of the two mediated interfaces. A range of behavioral measures were calculated from live measurement and analysis of the session recordings (Sec. 3.6).

3.2 Tasks

After a short warm-up task (a version of the Desert Survival Game, [37]), participants completed the following four tasks which were selected to cover different task types on McGrath's circumplex [39], which models different forms of group interaction. Task details and instructions are included in the appendix.

Estimation: An "intellective task" (Type 3 on McGrath's circumplex) involves solving problems with correct answers. Participants were asked to determine answers as a group to questions that require them to make statistical estimates. For example, "How many times does an average person blink in a day?"

Bribery: A "decision making task" (Type 4) involves coming to agreement on a matter that does not have a demonstrably correct answer. The experiment employed a moral judgment task in which participants act as a tribunal on a whistle blower case set in the workplace and must decide on appropriate punishment. A top salesperson has accepted an expensive paid trip from a client without reporting it. The salesperson's boss heard about this from a whistleblower, but failed to report it. The group needed to decide on a punishment for both while considering various stakeholders within the company.

Party Planning: A "Mixed-motive Negotiation Task" (Type 6) involves people coming to agreement when participants have different motives. Both mixed-motive tasks used a form of multi-issue bargaining, in which participants must agree on several different issues [21]. The scenario required the participants to agree on terms for a company party: the number of security guards to hire, the end time of the party, the price for guest tickets and how many knife jugglers to hire for entertainment. One participant was made Head of Security, one Head of Finance and the third Head of Social Planning, giving them conflicting interests. Each participant has a different points-table reward structure based on how each issue is settled, with conflicting and complimentary goals. They had time to study this before the session.

Floor Plan: The final task was also a mixed-motive negotiation task, but it introduced an artifact - a floor plan - to visually ground the discussion. Participants were told they are roommates that will be sharing an apartment. They need to decide on the room allocation and how to split the rent, with conflicting and complementary desires for rooms and additional features such as an extra closet. These were represented in a points table. In VR, the floor plan was placed on the table in front of participants. In VC, the floor plan was displayed on the same monitor with the remote participants and each participant had a different colored mouse they could use to point to items on the floor plan. This allowed a form of gesturing in both media to maintain an equivalent task.

3.3 Procedure

Upon arrival, participants were kept physically separated to ensure that all interactions between participants only occurred through the mediated interface. The participants were lead to three separate but similarly arranged areas (small conference rooms for VC, and partitioned areas for VR). To familiarize participants with the system and each other, the warm-up task was completed first, followed in randomized order by the four experiment tasks. Instructions were provided at the start of each task. Participants were told they could receive a reward of up to $11 per task based on their performance to incentivize engagement. All participants were paid the full reward at the end of the experiment.Each task could last up to 15 minutes, and the total time for an entire session was up to 195 minutes. Following each task, the remote connection was stopped temporarily for the participants to complete post-task surveys and to take a brief break. The experimenter's role was only to initiate the start and end of each task and to answer questions regarding task instructions.

3.4 Participants

There were 70 groups of three, a total of 210 participants, divided evenly between the two media conditions. Five participants did not provide demographic information. The remaining 205 were diverse in terms of age (M = 32.82, SD = 8.07), race (34 Black/African American, 39 Asian/Asian American, 4 Pacific Islander/Native Hawaiin, 84 White/Caucasian, 29 Latin/Hispanic, 2 American Indian/ Alaska Native, and 6 Other/Prefer not to say), gender (96 females, 105 males, 2 non-binary/third gender, 2 other), and education (15 completed high school, 48 completed some college, 23 with an Associate's degree, 93 with a Bachelor's, and 26 participants with a graduate degree). VR participants were diverse in terms of their experience in VR (35 with no prior experience, 43 with some experience with VR, 14 who had experienced VR several times, and 3 who own their own VR headset. 10 did not state their experience with VR).

Participants were recruited from a community participant pool and a research recruitment vendor. They were paid $200 for the 3 hour and 15 minute session, plus the $44 bonus. All participants were 18-45, without eye conditions that would impact tracking and comfortable speaking English. All groups were strangers, except one VC group that had two distant acquaintances. An effort was made to gender balance across media. The gender composition for video conferencing was 6 all-female groups, 7 all-male groups, 19 mixed groups, and 3 groups where we do not have enough information to classify. For the virtual reality condition, we had 8 all-female groups, 8 all-male groups, 16 mixed groups, and 3 groups where we do not have enough information to classify. One VC session was dropped due to failed eye tracking and one VR condition due to failed data recording.

3.5 Apparatus

Each participant used one of three identical *stations* featuring one of the two mediation interfaces:

3.5.1 VC. Each VC setup consisted of a participant looking at the other two participants on a 55" screen at a distance of roughly 5.5 feet. We used Zoom for videoconferencing with settings to keep participants in the same sized window throughout the task and their self view turned off. We followed the camera placement recommendations of Chen [13]. Eye Tracking was done with Tobii Pro Nanos. We used Microsoft LifeCam for the participant video streams. Interactions were recorded using OBS desktop capture. In addition, the video stream with the gaze overlay data from the Tobii Pro Software was recorded for later analysis.

Each participant saw two other participants horizontally laid out on screen for the Estimation and Bribery tasks. For mixed motive negotiations, the right half of the screen was vertically divided to show participants. The left half showed shared visual artifacts (the floor plan) and/or private points tables. A software called "UseTogether" allowed participants to each move a mouse on a shared floor plan.

3.5.2 VR. Participants wore a modified Oculus Rift Head Mounted Display (HMD) to view the VR scene of an office meeting room. They had their body, finger and face movement tracked in order to project them into the scene as avatars. Body tracking was performed using a single Kinect sensor as input and a custom motion solver that estimated the participant's skeleton pose. Body landmarks are inferred and an IK algorithm solves for skeleton joint angles by minimizing the squared distance of the observed landmarks and the attachment positions on the skeleton. Finger pose was calculated using HMD mounted cameras and a custom solver based on [24]. Face tracking was performed with cameras placed inside and outside of the HMD to view the participant's eyes and mouth. These cameras provide direct gaze tracking and were also used to estimate gains for a set of facial blendshapes in order to track facial deformation. Each station used two computers that

each housed a 12 core Intel Xeon processor, with 64 GB memory and either two or three Nvidia GTX 2080 graphics cards to perform tracking and stream combined motion into the VR scene. Overall, this provides direct tracking (not "head and hand" tracking) for a >100 DOF skeleton along with a 70 blend shape facial model, yielding very nuanced nonverbal behavior. All data is broadcast to a local network shared by the three setups. The system runs between 55-75fps.

A custom set of 36 avatars was built that included 3 male and 3 female avatars for each of 6 racial groups (Caucasian, East Asian, South Asian, African, Middle Eastern and Hispanic). This allows for basic matching between participants and their avatars. We ended up using 31 of those during the VR study sessions. During testing, it was found that extreme facial expressions on some avatars would create distracting artifacts, such as the eyeball penetrating the eyelid. To avoid these ever appearing during interaction, the range of the avatar expressions was reduced. A side effect of this was that overall expressiveness of facial motion was reduced. While speech activity was clearly visible on the animation of the mouth and lips, facial expressions were present, but damped.

3.6 Measures

Behavioral Measures were tabulated for conversational turn-taking (described in detail in Sec. 4), gaze (Sec. 5) and gesture (Sec. 6) behaviour. The analysis of the gaze data from the Tobii eye trackers in VC and internal HMD cameras in VR was largely automated and is described in detail in the Appendix. The other behavioral measures rely on annotations of the video logs of the sessions (Fig. 1). Videos were divided into one minute segments for annotation and annotators coded a single speaker at a time. Annotation was done by two remote annotation teams that were trained for this work. They were given detailed instructions that were then reviewed together with the research team. During training for both gestures and conversational turn annotation, annotators were given examples of correct annotations from the researchers. The final task for training was for each annotator to complete 20 annotations and for a researcher to manually evaluate and approve that they completed training successfully. During an initial test phase, annotators annotated clips for which a gold standard annotation had already been produced to check quality. After any issues were addressed, annotators proceeded to the main data. When they had questions on any part of the annotation, these were addressed by the research team. Annotations were spot-checked to ensure accuracy and annotators were encouraged to seek clarification throughout. Annotation was done using a customized annotation tool for remote video annotation.

Gesture annotation was completed by 4 annotators who were blind to the research goals and were not involved in data collection for the study. The annotators completed annotation using a predefined list of gestures and were asked to mark the start and stop of each gesture, the gesture label, and the reference label. The predefined list of gestures is detailed in the Appendix. Each 1-minute clip was independently annotated by 2 annotators, and if there was a mismatch between their annotations, then a 3rd annotator would review the annotations and arbitrate to produce a single annotation per gesture for our final analysis. Gesture annotation with a large label set is a challenging task given the subjective quality of co-verbal gesture – our annotators would initially agree on about 77% of gesture labels and 84% of reference labels – so the arbitration process provides a realistic method to achieve a high quality annotation. Spot checks were conducted daily by the researchers checking 20% of the annotations.

Conversational turn annotation was completed by 10 annotators who did not work on gesture annotation nor data collection, and were also not familiar with the research goals. Each 1-minute video clip was annotated once. Initial tests showed that this was a straightforward task that could be done accurately by a single person. Spot checks were performed by both an annotation manager and the researcher daily. After annotators completed annotations, the annotation manager reviewed 20% of their jobs and had any issues addressed. Once the annotation manager's checks were completed,

the researcher reviewed about 10% of all jobs submitted for each day, which included both those that went through the annotation manager's review and those that did not.

3.6.1 Statistical Tests Used. To avoid cluttering the discussion, statistical methods will be summarized here. Distributions were checked for normality. When normal, a linear mixed-effects model was fit to the data using the lmer() function in R. Linear mixed effect models are used to predict the dependence of a response variable (i.e. the item being measured, such as gaze duration) on one or more covariates (e.g. the Medium). They include both fixed and random effect terms, where a repeatable factor, such as Medium, is fixed and a non-repeatable factor, such as participant, is modeled with a random-effect term. Further details and information on the lme4 package which implements lmer() and glmer() can be found in [5, 6]. For non-normal distributions, a generalized linear mixed-effects model was used with either the glmer() or glmmPQL() [62] function and a log normal or Gamma distribution, depending on the data, as these provide a more accurate fit of the data. Significance of main effects and interactions was calculated using Anova, which performs Wald tests. Post-hoc tests were performed using estimated marginal means (emmeans() [51, 53]) which can be used with mixed effect models to compute pairwise comparisons which applies the Tukey method for correction. In some cases, a Wilcoxon rank sum test with continuity correction was used to compare two non-normal distributions and Bonferroni correction was applied as needed.

4 RESULTS: CONVERSATIONAL TURNS

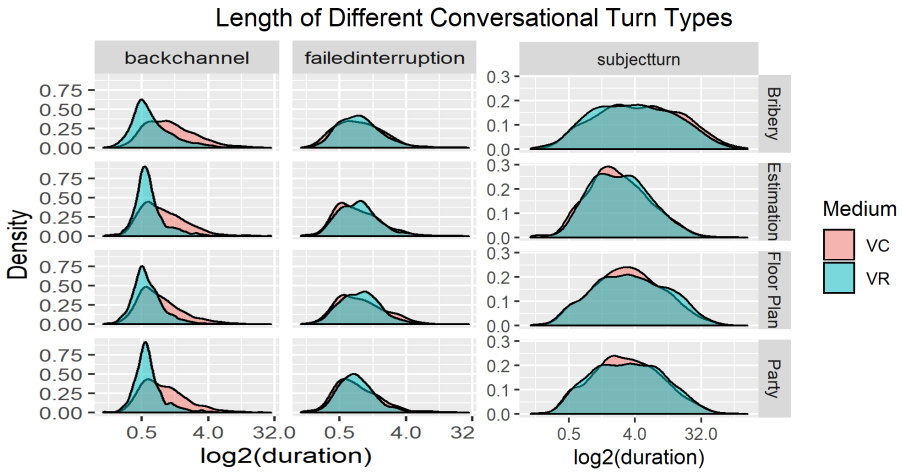

Fig. 2. Ratio of session time spent on all turns.

Analysis in this section focuses on three types of activities.

4.1 Turn Duration

A conversational turn (*speaking turn or subject turn*) is the period someone holds the floor while talking before yielding to another participant. Occasionally, nonverbal cues can be used to hold the turn, such as holding a hand out during a pause to indicate that you are not done speaking. Conversational turn length and interruptions provide an indication of the fluidity of conversation.

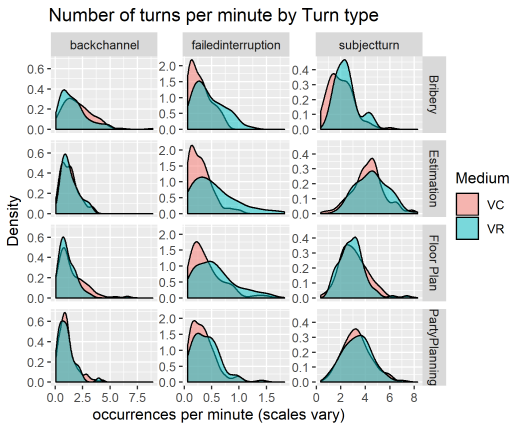

Fig. 3. Turns per minute.

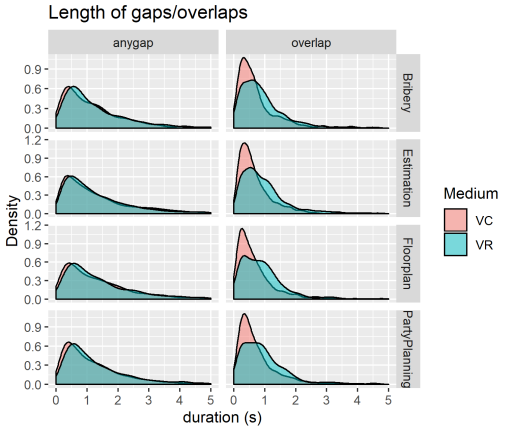

Fig. 4. Gaps and overlap.

For example, longer turns occur when people do not perceive that others want to speak[43, 64]. A *Backchannel* occurs when the listener provides acknowledgement, such as saying “mhmm" or nodding their head. *Failed interruptions* occur when someone tries to take the conversational floor, but the person speaking does not yield.

We find that the length of speaking turns (Figure 2) is impacted by task, with an ordering from longest of: Bribery (median 4.08s , mean 8.94s, sd 12.5), Floor Plan (median 3.24, mean 6.14, sd 7.71), Party Planning (median 3.09s, mean 5.63s, sd 7.24) and Estimation (median 2.28s, mean 3.88s, sd 4.75). In all cases, these differences are significant for both media at $p < .0001$ except for the differences between the Floor Plan and Party Planning tasks. The latter are significant for VR (t.ratio = 2.762, p= 0.029), but not VC (t.ratio = 2.24, p= 0.11).

The impact of medium is reflected in significant interactions. The largest difference is for Bribery, where VC turns are significantly longer, over 20% on average (t.ratio = 9.88 , p<.0001; VC median 4.41s, mean 9.86s, sd 13.8; VR median 3.73s, mean 8.13s, sd 11.04). Turns are significantly shorter for VC in Floor Plan (t.ratio = -2.46, p =0.014; VC median, 3.17s mean 5.76s, sd 7.22; VR median 3.38s, mean 6.53s, sd 8.19) and Party Planning (t.ratio = -2.023, p =0.043; VC median 3.00, mean 5.50, sd 7.20; VR median 3.18, mean 5.80, sd 7.29), but these differences are somewhat less marked, averaging 13 and 5 percent respectively. Durations were not significantly different for Estimation.

There is no significant difference in the length of failed interruptions.

The duration of backchannel turns is significantly longer for VC than VR (Chisq 738.45, Df 1, p < 2e-16; VC median 0.91s, mean 1.42s, sd 1.62; VR median 0.58s, mean 0.77s, sd 0.87) and this relationship remains significant across all Tasks. There is also an effect of Task (Chisq 278.71, Df 3, p< 2e-16). Post-hoc analysis reveals that backchannels are longer for Bribery than all other tasks and this relationship holds for both VC and VR (Bribery median 0.82s, mean 1.36s, sd 1.66; Estimation median 0.65s, mean 1.01s, sd 1.21; Floor Plan median 0.64s, mean 0.96s, sd 1.05; Party Planning median 0.68s, mean 1.08s, sd 1.64)

4.2 Turn Frequency

The duration of the turn provides one characterization of conversation. Frequency of turn types is an important complement (Figure 3). Examining backchannels per minute shows a significant main effect for Medium (Chisq 7.4957, Df 1, p = 0.0062) and Task (Chisq 71.2297, Df 3, p<.0001), but no significant interaction (Chisq 4.2161, Df 3, p = 0.24). VC backchannels are more frequent (VC median

1.18 per min., mean 1.49, sd 1.12; VR median 1.06, mean 1.30, sd .95). Post-hoc analysis shows that backchannels are more frequent in Bribery than all other tasks (all p<.0001). The frequency was also significantly higher in Estimation than Party Planning (z.ratio -3.184, p = 0.0079. The overall frequencies by Task are Bribery (median 1.60, mean 1.89, sd 1.33); Estimation (median 1.16, mean 1.33 sd .79); Floor Plan (median .98, mean 1.30, sd 1.00); Party Planning (median .92, mean 1.06, sd .78).

An analysis of the frequency of failed interruptions showed a significant main effect of Medium (Chisq 45.55, Df 1, p<.0001) and of Task (Chisq 13.75, Df 3, p=0.0033) and a tendential interaction (Chisq 6.85, Df 3, p=.077). Failed interruptions are more frequent in VR (VC median .29 per min., mean .33, sd .23; VR median .43, mean .49, sd .34). Post-hoc analysis suggests that they were more frequent in Floor Plan than Bribery (z.ratio 3.182, p=0.0080) and Party Planning (z.ratio -2.962, p=0.0161) (In order by mean occurrences per minute: Floor Plan median .42, mean .49, sd .34 Estimation median .36 mean .43, sd .34 Bribery median .31 mean .38, sd .25 Party Planning median .32, mean .37, sd .24.)

Turning to the frequency of speaking turns, there was no main effect for Medium (Chisq 2.0485, Df 1, p=0.15), but there was a main effect for Task (Chisq 300.15, Df 3, p < 2e-16) and a significant interaction (Chisq 10.7463, Df 3, p = 0.013). There is a significant interaction between Medium and Task for Bribery (z.ratio 3.332, p = 0.0009), with VC being less frequent (VC 2.02 vs. VR 2.42). The differences between all Tasks are significant for both Media, except Floor Plan and PartyPlanning, which are significant for neither (VC z.ratio=1.341, p=0.5370; VR z.ratio=2.496 p=0.061). Tasks by order are: {Bribery median 2.13, mean 2.21, sd 1.01}, {Floor Plan median 2.95, mean 2.99, sd 1.11, PartyPlanning median 3.28, mean 3.32, sd 1.16}, {Estimation median 4.48, mean 4.38, sd 1.31}.

4.3 Conversational Gaps and Overlap

We looked at both the length of the gap between speaker turns and the amount turns overlap (Fig. 4). The gap duration data did not show a consistent pattern, but there is a clear difference for overlap. There is only a significant main effect of Medium (chisq 104.64, Df 1, p<2e-16). This shows that overlaps are significantly longer in VR (VC median .52, mean .71, sd .62; VR .77, mean .93, sd .72).

4.4 Discussion

Bribery is clearly distinguished from the other tasks. It had the longest speaking turns, the longest backchannels, the least frequent speaking turns and the most frequent backchannels. *In short, people spoke longer but less frequently, and they both provided more backchannels and these had a longer average duration.* The Bribery task is social, subjective and collaborative. The longer speaking turns could be explained by the need to make more complex, and hence longer, arguments, given the subjective material. The more frequent and longer backchannels would seem to reflect the need for increased coordination as there was a need to reach consensus on a more subjective and emotional task. By contrast, during Estimation people were often sharing short facts or guesses, which could explain the Estimation speaking turns being the shortest and most frequent.

The most striking difference between media occurs around backchannels, which were both longer and more frequent in VC than VR. Backchannels serve to maintain the social connection, acknowledging that the other person is heard and understood. People felt a need to do more of this "connection maintenance" in VC. Another possible explanation is that people were simply less connected or more tuned out in VR, but social presence surveys run during the experiment showed similar levels across the media, speaking against this explanation.

Previous work has shown that turn taking in video conferencing (VC) is more formal than in face-to-face communication [20, 43]. It appears that it is also more formal in VC than embodied VR.

Failed interruptions were more frequent in VR and there were longer overlaps of speaking turns in VR. Both suggest less careful adherence to strict turn taking behavior. It is important to note that in our coding, failed interruptions included all times a person interrupted or interjected, but did not obtain the floor, so they are not necessarily negative. Listening to the sessions, it appears that people were more comfortable and more able to effectively talk over each other in VR by providing brief interjections of helpful content, whereas they followed a more strict turn taking approach in VC. This is also consistent with the increased overlap in VR.

One could argue that this overlap occurred because people receive less clear signals in VR that they are speaking over top of another and it took them longer to detect this "error". However, the fact that the duration of failed interruptions was similar across media speaks against this. It suggests both media provided sufficient clues that the interruption would not be successful (or neither media made it clear to the speaker that someone was trying to interrupt), and suggests that people gave up on attempts to interrupt after a similar effort. People appear to be more comfortable overlapping dialogue in VR.

The length of speaking turns is longer in VC for Bribery, but shorter in VC for Floor Plan and Party Planning. For the latter two tasks, the payoff tables in VR were displayed with a virtual touch interface that users found a bit difficult to use, which might have led to longer turns. The differences are much larger in Bribery. Shorter turns tend to occur when there are more clear nonverbal signals for turn taking [64], which might explain the difference in Bribery, but conclusions should be drawn with caution as VR only outperformed in this one task.

5 RESULTS: GAZE

Technical details on gaze tracking analysis are contained in the appendix.

5.1 Categories of Gaze

A first analysis considers the distribution of gaze by task and condition. Gaze is broken into three broad categories: *Body,* which includes gaze at any other participant; *Task,* which includes gaze at task artifacts when they exist (the payoff tables or floor plan); and *Elsewhere* which includes all remaining gaze.

Figure 5 shows the portion of time participants looked at various body parts in VR and VC. This data was best fit to a generalized linear mixed effects model with a Gamma distribution, containing all main effects and interactions. Type II Wald chisquare tests show no main effect for Medium ($\chi^2(1) = .0271$, $p = 0.8692$), but significant main effects for Task ($\chi^2(3) = 1241.9$, $p < 2.2e-16$) and Category ($\chi^2(2) = 1471.0$, $p < 2.2e-16$), as well as all interactions being significant: Medium:Task ($\chi^2(3) = 33.37$, $p = 2.697e-07$), Medium:Category ($\chi^2(2) = 344.2$, $p < 2.2e-16$), Task:Category ($\chi^2(4) = 148.4$, $p < 2.2e-16$), Medium:Task:Category ($\chi^2(4) = 208.4$, $p < 2.2e-16$) . Table 1 shows the means for VC and VR in each Task and gaze target category, along with post-hoc statistics computed using the Tukey method and emmeans(). Participants in VC spent significantly more time looking at their interaction partners than in VR across all tasks, on average, *56% more time.* Participants in VR spent significantly more time looking at Task artifacts than in VC for the Party Planning task.

Ordering by the mean proportion of time, people spent the most time looking at other participants in Bribery, followed by Estimation, Party Planning and Floor Plan. The difference between Task was significant in all cases except the difference between Party Planning and Floor Plan in VR.

5.2 Gaze and Body Areas

The time participants spent looking at another participant is broken down by the areas of the body they gazed at in Figure 6. If participants only briefly gazed at another participant, these ratios may

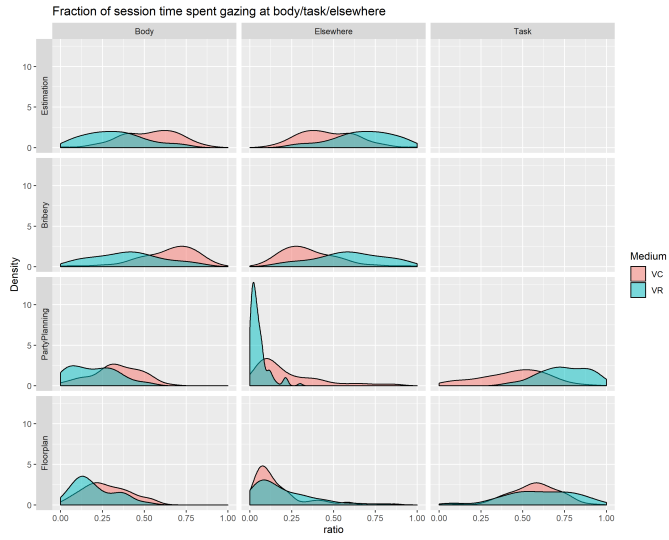

Fig. 5. Ratio of gaze at different targets across Task and Medium.

	Body				Elsewhere				Task Item			
Task	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p
Estimation	.532	.325	-8.429	<.0001	.468	.675	7.601	<.0001	-	-	-	-
Bribery	.642	.406	-8.973	<.0001	.358	.594	9.173	<.0001	-	-	-	-
PartyPlanning	.334	.201	-6.216	<.0001	.222	.0517	-8.875	<.0001	.444	.747	11.210	<.0001
Floor Plan	.275	.202	-3.366	0.0008	.162	.197	1.725	0.0846	.563	.600	1.344	0.1789

Table 1. Proportion of time spent looking at other participants (Body), at something not task related (Elsewhere) or a task artifact. When differences are significant by medium, they are color coded pale red for the more frequent, blue for less.

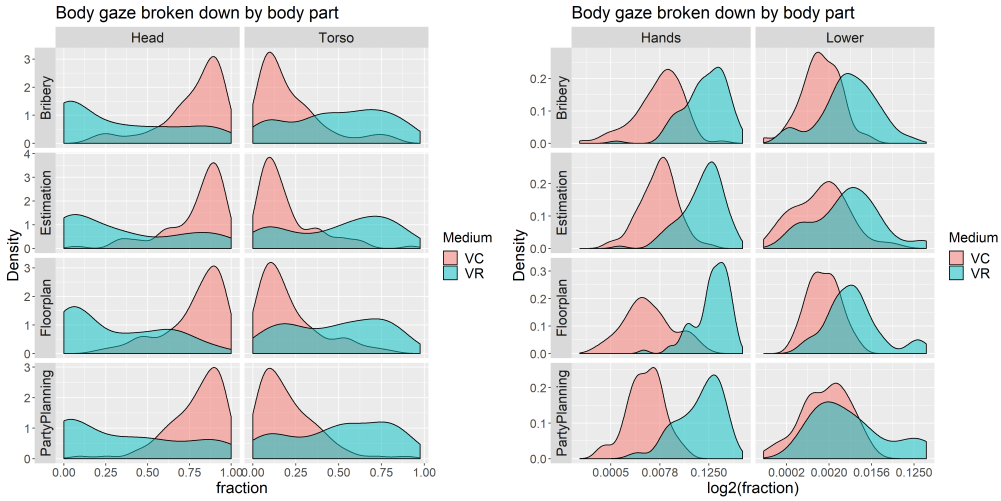

Fig. 6. These figures show the distribution of time looking at different body parts when a participant is gazing at the body. Note that the Head and Torso figure on the left is on a linear scale and the Hand and Lower body figure on the right is on a log scale

	Head				Torso				Hands				Lower Body			
Task	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p	VC	VR	z ratio	p
All	.78	.34	-27.07	***	.21	.49	20.42	***	.010	.16	13.43	***	.0013	0.0049	0.35	0.73

Table 2. Proportion of time spent looking at different body parts, averaged across Task because Task did not lead to significant variation. The table shows means for VC and VR. When differences are significant by medium, they are color coded pale red for the more frequent, blue for less. *** indicates p<.0001

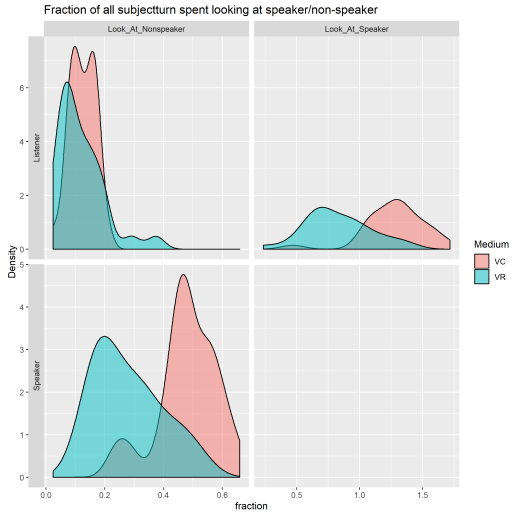

Fig. 7. The proportion of time people spend looking at speakers and listeners during conversational turns.

not have stabilized. Therefore any *participant sessions* (a single participant on a single task) for which there were less than 45 seconds of body gaze were dropped from the analysis. This left 677 participant sessions for analysis.

Models were fit to the data by progressively adding the factors Category, Medium and Task. The best model fit was obtained using a Laplace approximation for a generalized linear mixed effect model with a Gamma distribution for gaze Category and Medium. This indicates that Task did not have a significant effect on gaze distribution, which is consistent with Figure 6. There was a significant main effect for gaze Category ($\chi^2(3) = 4903.3658$, $p < 2e-16$), but not Medium ($\chi^2(1) = 0.12$, $p = 0.73$). There was a significant interaction between Body and Medium ($\chi^2(3) = 1329.88$, $p < 2e-16$). The means and significant differences for Medium are summarized in Table 2, which shows that Medium had a significant impact on every Category but Lower body. Differences between Categories were always significant except for the difference between Hands and Lower Body for VC.

Head and Torso are the main gaze targets overall. For VC, the Head is by far the dominant focus of attention (78%), with the Torso receiving the bulk of remaining attention (20%). Participant gaze in VR is more dispersed across the body. The Torso was the most significant category (49.2%), followed by Head (34.3%) and Hands emerge as an important target (16.0%). The Lower body receives little direct attention in either medium. It was also largely obscured by the edge of the screen in VC and table in VR. It is worth noting that these statistics only account for direct gaze. Participants may of course still perceive movements of other body parts through their peripheral vision.

5.3 Gaze During Conversational Turns

Figure 7 shows how frequently participants look at the speaker or listeners during conversational turns. Note that for one speaker, there are two listeners and this figure sums results over the total number of people in a role (i.e. if both listeners look at the speaker throughout, this would produce 200% listener-at-speaker gaze). Since the data is not balanced and there is no anticipated relationship between the categories, separate tests were used to compare the impact of Medium on each condition. A speaker gazed at a listener an average of 48.1% of the time in VC and 28.0% in VR. The distributions are significantly different according to a Welch Two Sample t-test (t = 7.0329, df = 54.843, p = 3.38e-09). The listeners gazed at a speaker on average 128.4% of the time in VC and 82.5% of the time in VR. This difference is statistically significant according to a Welch Two Sample t-test (t = 6.9751, df = 55.547, p = 3.95e-09). Listeners would gaze at non-speakers on average 12.4% of the time in VC and 11.9% of the time in VR. Since the VR distribution is not normal, a t-test based on various robust estimators was used and indicated no significant difference (pb2gen, Test statistic: 0.0211, p = 0.27).

5.4 Discussion

The two media are clearly distinguished by differences in gaze behavior. People spend much more time looking at each other in VC than VR, on average 56% more time, and when people do look at each other, they are much more likely to look at the head in VC (78%) than in VR (49%), where gaze is more distributed, with the Torso (34%) and Hands (16%) also receiving notable attention. This behavior occurs for both speakers and listeners. Speakers in VC look at a listener 48% of the time, compared to only 28% in VR. Listeners look at speakers on average 64% of the time in VC versus 41% of the time in VR. Interestingly, listeners look at other listeners about the same in either medium. This seems to suggest that the increased gaze is therefore about the speaker-listener interaction or connection, and not about maintaining a connection to everyone in the triad.

We postulate that this additional gaze felt necessary in order to maintain the social connection in VC. This is based on a reflection on the many functions of gaze. One function is to both give and show attention [34]. People are clearly giving more attention to the people at the other end of the speaker-listener interaction in VC. This could reflect that people are for some reason more interesting in VC, perhaps because their facial features are rendered with higher fidelity. If this was so, however, it seems reasonable that there would be some difference reflected in subjective measures of social presence, and those were not seen. Conversely, it is possible that the VR environment was more interesting, but both the VR and the VC environments were quite plain offices. Perhaps a more likely explanation is that this increased attention feels necessary to maintain connection with someone who is remote and located in a different 3D space. This need could be driven by two sources: people might feel that they need to look at the other person so that they are not distracted by other items in their own space which the remote participant cannot in general see or perhaps they feel a need to show to the other person that they are paying attention by directing their gaze at them. We expect the latter may be dominant, but this requires further investigation. When people are co-located in the same space in embodied VR, a lower level of gaze may have felt sufficient to maintain the connection.

Another function of gaze is to regulate turn taking [64]. If other methods of turn taking regulation, such as gesture, are less effective in VC, this might lead to increased reliance on gaze cues, and hence increased gaze, although Beattie suggests gaze becomes a less effective turn taking cue as the baseline level of gaze increases [8].

There are also costs associated with gaze. It is a powerful cue [64] that can express intimacy and exercise social control [34]. It is known that there is social pressure associated with receiving

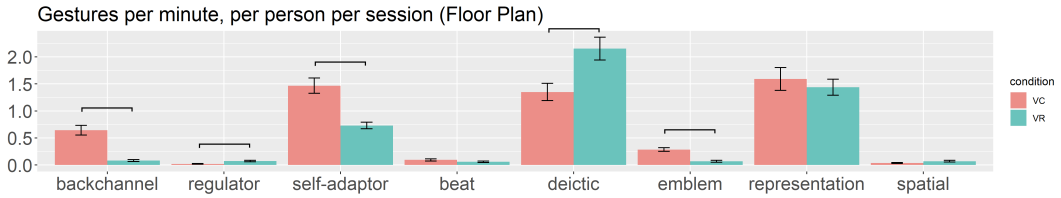

Fig. 8. A comparison of gesture rates between VC and embodied VR for the visual aided out FloorPlan task. Significant differences are marked.

directed gaze, and this has even been exploited in VR by artificially increasing gaze to increase listener attention [2, 3]. There may also be costs with *giving* gaze. It is worth considering what burden this increased gaze may place on users.

Explanations for the more distributed attention in VR could include that people gathered more information from other body parts in VR, perhaps because gesture was more effective, or again that they felt more of a need in VC to look directly at a speaker as a way to show attention. It may also be that the face in VR provided a less reliable signal with the current state of technology, so people relied more on other signals. Another possibility is that the people in VC occupied less of people's field of view, so they could see the whole person by staring at their head, but gaze needed to be more distributed in VR to fully observe the person.

Increased gaze during Bribery could occur because this is the most subjective, social task and so relies heavily on people being able to read social signals. Party Planning and Floor Plan both featured Task objects that garnered significant attention. They may have looked at these frequently to remember the payoffs or to intentionally avoid gaze in a competitive task. The increased task gaze in VR Party Planning may be due to delay caused by the increased difficulty of using a virtual touch interface compared with a mouse.

6 RESULTS: NONVERBAL ANALYSIS

A detailed annotation (details in Appendix) was completed of the gesture behavior for participants in the Floor Plan task. Due to the high effort involved in such an annotation, we were only able to complete this for a single task and the Floor Plan task offers the richest visual grounding, so is particularly appropriate for a gesture analysis.

6.1 Gesture Frequency

Figure 8 shows the frequency of different gesture types across media. Each gesture type was analyzed separately to see if it varied based on Medium. Applying Bonferroni correction on eight tests yields an $\alpha = .00625$ threshold. Since the distributions are not normal, Wilcoxon rank sum tests are used throughout.

Backchannels are significantly more frequent in videoconference ($W = 5723.5$, $p < .0001$, VC median .30 per minute, mean .64, sd .81; VR median 0, mean .083, sd .18). Self-adaptors are also significantly more frequent in videoconference ($W = 4882$, p-value = 0.00015, VC median 1.19, mean 1.47,sd 1.29; VR median .57, mean .73, sd .58). Regulators, gestures that explicitly turn over the conversational turn, were quite rare with only 73 occurrences in the corpus, but significantly more common in VR ($W = 2860.5$, p-value = 0.0018; VC median 0, mean .021, sd .049; VR median 0, mean .075, sd .13).

There was not a significant difference in the rate of Beats ($W = 4182$, p-value = 0.048). Deictics were significantly more frequent in VR ($W = 2644$, p-value = 0.0018; VC median .88, mean 1.35,

sd 1.46; VR median 1.74, mean 2.15, sd 1.96). Emblems were uncommon, but significantly more common in videoconference (W = 5281, p-value = 4.5e-08; VC median .21, mean .29, sd .31; VR median 0, mean .071, sd .16). The rate of Representational gestures did not vary significantly based on medium (W = 3469, p-value = 0.57). Spatial gestures were also relatively uncommon and did not differ significantly by Medium (W = 3365, p-value = 0.26).

In terms of novel information in gestures, 5.3% of gestures were marked reference pronoun (definitions in Appendix, Table 4) in videoconference compared to 15.8% in VR. Other unique information occurred with 6.65% of gestures in videoconference and 4.80% in VR. Overall, 12.0% of videoconference gestures and 20.6% of VR gestures contained unique information. All these differences are significant using a 2-sample test for equality of proportions with continuity correction ($\chi^2(1) = 220.24$, $p =< 2.2e-16$, $\chi^2(1) = 11.955$, $p = 0.00054$; $\chi^2(1) = 103.5$, $p < 2.2e-16$)

6.2 Discussion

As with conversational turns, nonverbal-only backchannels are more frequent in videoconference. *Assuming that people are not more agreeable in one condition than the other, this may show a greater felt need to actively maintain the social connection in videoconference and is consistent with behavioral changes seen for conversational turns and gaze.* Self-adaptors are associated with anxiety [63], so their increased prevalence in videoconference could suggest less comfort.

One explanation for increased person-directed gaze in VC is that this is required for turn regulation. The increased use of regulators in VR could reflect the greater ease of passing the turn in a 3D environmentusing gesture, compared with having two interlocutors on a screen, which would be consistent with the discussion related to gaze. However, these were still rare events, so do not suggest a major change in behavior between media, and hence, turn regulation is unlikely the dominant reason for gaze differences. There are also multiple syntactic and paralinguistic cues for turn taking that could be deployed [8].

Even though every effort was made to provide a shared mouse interface that facilitated pointing, it is still likely easier to point in a shared 3D environment, so it is not surprising that deictics were higher in VR. The increased emblem use in videoconference could be because people found them easier to use in this mediumif hand tracking was not perfect, gestures like deictics were less accessible so people switched to emblems or perhaps there was more use of “thumbs up" and similar gestures in an effort to maintain a social connection. Again, these were relatively infrequent, so strong inferences should not be drawn.

There was a marked overall increase in unique information conveyed by gestures in the VR condition, although videoconference had more unique information that was not associated with a reference pronoun. The increase in reference pronoun use in VR exceeds the increase in deictic gestures alone. This likely suggests that it was more fluid in VR to point at things, so people were more likely to use that pointing as a substitute for speech.

7 DISCUSSION

This section will relate our work to previous research and offer interpretations for our findings.

Task: Considering the Task factor first, *the behavior during Bribery is clearly differentiated from the other tasks. It had the longest speaking turns, the longest and most frequent backchannels, and the most person-directed gaze. The Bribery task requires a high level of coordination between the participants.* It is a subjective task where the group must strive for mutual understanding and to reach consensus. Estimation also requires consensus, but on factual questions so it lacks the subjective nature of Bribery. Both Floor Plan and Party Planning are negotiations that can be viewed, at least in part, as adversarial and do not involve a sensitive discussion of moral norms nor group consensus, like Bribery. The longer turns in Bribery may suggest more elaborate arguments.

Longer backchannels and increased gaze seem to reflect attentive listening, which is consistent with the task. This increase in prosocial behavior provides evidence related to theoretical arguments such as information richness theory, that suggest that richer (more embodied) media are better when equivocality is higher [15] and there are needs for group coordination [61]. People are at least making more use of additional affordances on the most equivocal task, whether or not this actually improves their performance. The increased gaze and backchannel behavior is also consistent with research suggesting embodiment is more important for social, interpersonal communication, whereas speech is more task-oriented [64].

Performance: Turning to the impact of Medium, while our detailed performance analysis is not included in this paper, *it is perhaps unsurprising that we did not observe performance differences between VR and VC.* Much previous work comparing audio and video interfaces has failed to show a critical impact of visibility on completion time or quality of work [20] and given that audio lacks any visual component, it is arguably a more different interface than the two visual media explored here. Evidence for the benefit of video tend to be for physical manipulation tasks [20], which were not explored here. Even comparing text and face-to-face, Strauss and McGrath [61] showed similar quality output across the two interfaces, but face-to-face was more efficient and participants were more negative towards using text for their judgment task, which is similar to our bribery task, a difference we did not see for the media in this study.

Considering the limited related work on avatars, Bente et al. [9] also did not see notable performance differences between videoconference and avatar based interaction. Jo et al. [31] found that video performed worse than avatars on a job interview task, but for a measure that combined spatial and social presence questions, making it difficult to interpret. *Our contribution to this literature is to show that there are strong behavioral differences between embodied avatars and videoconference, even if performance is similar.*

Conversational Turns: Previous work based on conversational game analysis found that face-to-face interaction was more efficient than audio-only because participants would rely on visual indicators to confirm understanding, rather than requiring verbal feedback, but this efficiency improvement did not hold for video when compared to audio [17]. *A key finding of our work is that backchannels are longer and more frequent in VC than VR. This appears to reflect a greater need to align and affirm in the medium*, as was seen with audio and VC in the past.

The more formal nature of turn taking in audio and VC has been discussed above. For example, one study found people interrupted each other around twice as much in face-to-face as VC (12.6 to 6.5 interruptions per dialogue) [17]. *Our work shows a less formal interaction style in embodied VR, with more failed interruptions and more overlapping speech. It thus appears more similar to face-to-face.*

Gaze: *Previous work found the same increase in gaze for VC as compared to face-to-face that we observed for VC compared to VR.* A central finding of this paper is that people gaze at each other much more in VC than embodied VR, and this gaze is more focused on the head. Previous work has found higher rates of gaze in VC than face-to-face [11, 17, 44]. The study reported on in O'Malley et al. [44] and Doherty-Sneddon et al. [17] shows that subjects using video gazed at each other 56% more than subjects who were face-to-face. This is exactly the same increase that we saw for gaze in VC vs. embodied VR, suggesting VR is more in line with face-to-face. They refer to the VC behavior as "overgaze," since it represents an unnatural increase over the face-to-face baseline. Interestingly, they found co-present subjects would gaze much more when speaking than listening, but there were no such differences in video [44], whereas we found listeners gazed more at the speaker in both media, as was also reported in recent research on in person conversation [38]. Setlock et al. [55] reported much lower gaze levels in a video-only study that used the desert survival game as a task, our warm up exercise. They found gaze never occurred more than 25% of the time, whereas

our mean for VC was 45% for speakers gazing at listeners and 64% for listeners gazing at speakers. It may be that this decrease in gaze occurred because they gave participants a paper printout of the desert survival items which they could place in front of them during the exchange and they felt free to look at. Other differences may also be important, such as the closeness of the screen or changes in use of VC that may have occurred as it became prevalent. Finally, we studied triads and group size has been shown to impact conversational gaze [38]

Gesture: *Several findings suggest nonverbal behavior may have been more effective in VR, especially when it had a spatial component.* Participants had longer conversational turns in VC for Bribery, which could suggest they were less aware of attempts to take the floor. This effect was reduced but opposite for Floor Plan and Party Planning. Both of these tasks featured a more difficult to operate virtual touch interface, so it is difficult to say if these cases offer counter evidence or the duration is related to people trying to display their points table. Deictic gestures were much more frequent in VR and people were much more likely to replace a word with a reference, likely suggesting that pointing was easier. Previous work has suggested that deictic gestures may increase efficiency of communication [64], but we saw no clear evidence of that. Other forms of unique gesture content were slightly higher in VC, but overall, information that was only available in gesture was much higher in VR (20.6% of gestures compared with 12% for VC). Regulators that manage the conversational turn and are often directed towards a speaker were uncommon, but more likely in VR. Emblematics were also uncommon, but show the opposite trend, being more likely in VC. They generally have no spatial information.

Interpretation: *People were making more effort in VC, in terms of increased prosocial behavior like backchannels and gaze, in order to maintain a similar level of connection as reported on the subjective surveys.* The subjective surveys conducted after each task suggest that participants were experiencing a high level of social connection in both media, and most importantly, there were not notable differences between VR and VC. It is therefore difficult to attribute the observed behavioral changes to differences in social experiences of the media and other explanations must be sought.

To understand what might be underlying the behavioral changes, it is helpful to clarify the differences between our VC and embodied VR manifestations by comparing them on the factors of the decompositional model of copresence developed by Kraut et al. [36] and summarized in [20]. *Field of view* relates to whether people can see what entities other people are oriented towards. This is easier in our VR setup as people can freely adjust their orientation and other interlocutors can tell what they are oriented towards because they can see the same items. The *spatial perspective* is similar in both media as participants are seated and hence can see their own view, but cannot reorient to share their interlocutors perspective (something that would be easy in VR in general, but we intentionally restricted). *Display symmetry* differs as in VR, interlocutors can all see essentially the same environment, but in VC, much of people's local environment is off camera and not visible to their remote interlocutors. The *dimensionality* is 3D in VR and 2D in VC. *Spatial resolution* is arguably higher in VC as people's facial details are rendered with more fidelity. *Temporal delay* is low and comparable in both. *In short, the critical differences between media arise from users being in a shared, 3D environment in VR, combined with a secondary factor of arguably higher visual fidelity of VC.*

We postulate that the behavioral differences across media arose because people are making a greater effort in VC in order to actively manage the social connection. Gaze performs multiple social functions in conversations, including providing information such as cues for attention and competence, regulating the interaction, expressing intimacy and exercising social control[34]. Higher gaze leads to increased compliance [54] and gaze aversion has shown consistently negative effects [12]. While these findings suggest why people might feel the need to maintain gaze in general, they do not address the differential between VC and VR. *Given that subjective experience*

was similar, we postulate that these behavioral differences arise from the main difference in the media: that VR places people in a shared, 3D space. People's substantially increased tendency to look at each other in VC could reflect that such visual attention was needed for them to remain engaged, or perhaps more likely, they fear that not showing such attention would be viewed as rude or inattentive. The gaze differences may arise from people wanting to show that they are in the same environment as their interlocutors, rather than looking at items only they can see in their local environment. This is not an issue in VR as people are de facto in the same virtual space.

Further evidence is consistent with the hypothesis that people felt a greater need to actively maintain the social connection in VC. Backchannels are a social feedback mechanism, so their increased use in VC suggests effort to manage the connection. This connection maintenance explanation is also consistent with Bailenson's hypothesis for explaining Zoom fatigue whereby he suggested people experience increased cognitive load in Zoom because of increased effort to send and receive nonverbal signals, including attention [1]. Zoom often creates increased intimacy by placing people close to a large image of another's face [1]. Interestingly, we intentionally took a wider viewing angle and greater distance, giving people more latitude to look elsewhere on the screen, but they *chose* to look at each others' face a high proportion of the time. By contrast, people looked at each other less in VR, and when they did, gaze was more distributed between the head, torso and hands. Finally, it is worth noting that people displayed more self adaptors in VC, which are related to anxiety [63], possibly suggesting less comfort, which could actually be triggered by the increased gaze and/or effort to maintain the connection.

It is important to also consider other potential explanations. In trying to explain similar gaze differences, O'Malley et al. "suggest that when speakers are not physically co-present they are less confident in general that they have mutual understanding, even though they can see their interlocutors, and therefore over-compensate by increasing the level of both verbal and nonverbal information" [44]. They also suggest that to gain as much information as they would in face-to-face interaction, more gaze may be necessary, or that people felt their communication was less effective, so compensated with more gaze [17]. These are also plausible explanations for the differences seen with embodied VR, worth considering, although their exact mechanism of action is unclear. They can be contrasted with the much lower gaze rates observed in Setlock et al. [55], which may be more easily explained within the framework of a social connection hypothesis. The gaze reduction could have occurred because both participants were given a paper artifact by the experimenter and it was clearly part of the task, so they did not feel rude looking at it and it felt socially acceptable to reduce gaze. If gaze was needed for understanding, it is less clear why it would have been lower. Explanations of interface novelty that were offered in the past to possibly explain increased gaze in VC [17] seem less likely now that it has become a common form of interaction, and embodied VR is likely more novel for all participants.

Although we did not observe performance differences, there may be negative consequences of these behavioral changes. O'Malley et al. [44] raise the concern that "overgaze" may increase cognitive load and inhibit verbal processing, since people tend to avoid gaze during periods of increased cognitive load [7]. Bailenson raises concerns about fatigue in videoconferencing [1]. Beattie [8] argues that evidence from face to face studies suggest that gaze may be most effective for turn management when its background level is low, so it may become a less effective cue in VC.

When considering overlapping speech, failed interruptions and gaze, the behavioral differences observed between the media indicate that embodied VR is closer to what has been observed for face-to-face interaction [17, 44] and suggest less formal interaction.

Relying on a definition from Clark and Brennan [14], Fussell and Setlock [20] view co-presence as sharing the same physical environment and argue that it is particularly important for collaborative physical tasks, such as repairing an item. *Our work shows that co-presence, as achieved by being in a*

shared 3D environment in VR, impacts behavioral patterns even when people are not working on such physical manipulation tasks.

There are also additional factors worth considering. Although people were told their avatar looked like them, they could see the avatars of other people were of moderate fidelity, so this likely provided some sense of anonymity compared with video. This could have influenced their interaction. Facial motion was reduced due to limitations in the avatars, but they did exhibit facial motion, body motion and finger motion that tracked that of their users, likely providing one of the highest quality avatar experiences in a study of this scale. To explore an effort hypothesis, it would also be valuable in follow up work to measure energy expenditure or similar qualities to see if the behavioral changes took a toll on participants.

8 CONCLUSION

This paper discusses a study comparing behavior during remote, multiparty meetings across two conditions. The first condition was Medium. Participants interacted through one of videoconference or embodied virtual reality. A second, within subject condition was Task type, that included four different tasks that match the types of activities people often perform during meetings: an intellective task (Estimation), a decision making task where groups needed to reach consensus on a problem without a clearly correct answer (Bribery), and two mixed motive negotiation tasks, one with visual grounding of the conversation on a map (Floor Plan) and one without (Party Planning). *Notable behavioral differences were observed. People showed longer conversational turns, more gaze and more backchannels for the Bribery task that required a high level of coordination and subjective discussion. Interaction in VC showed more frequent and longer backchannels, more formal turn taking, much higher rates of gaze, more gaze at the face, more self adaptors and less unique information in gestures, among other differences. In comparison with previous work, the levels of VR behaviors often appear closer to face-to-face interaction. To explain some of these differences, we postulate that people needed to make increased effort in VC to maintain a similar level of social connection.*

We feel that embodied VR has great potential to be a beneficial technology for remote collaborations. We find it encouraging that some of the behavioral patterns observed here are closer to what has been observed in face-to-face interaction than those produced through VC interactions. While much more work is needed to understand the impact of these behavioral differences, and certainly VR is a long way from matching the richness of face-to-face interaction, we hope that this provides early evidence of the potential of VR to support more natural remote interactions. Reducing the environmental impact of travel alone makes this a worthwhile goal.

In closing, it is worth noting that VR is a far more flexible technology than was explored here. For the purpose of this study, people were required to remain seated in our Embodied VR condition to maintain equivalence with VC, as well as replicating a common meeting scenario, but this is in no way a technological requirement. One of the strengths of VR is that people can walk around in their space, they have full control of their viewing direction, and they can potentially even group up and have side conversations, as people do in real life. There is evidence that sharing a common viewpoint is particularly useful in supporting physical manipulation tasks [20] and we would like to explore the appropriateness of VR for such applications in the future. In addition, it is important to undertake longitudinal studies of extended work collaborations conducted through VR, these would include both much longer individual sessions and work that continued over days or months. VR also offers great potential for visual and aural displays than was investigated here. People could start a meeting around a conference table, transfer to a construction site and then switch to a design studio, all while remaining in the same physical space. Much work is needed to understand how to

make best use of these affordances. Finally, it would be interesting in future work to see if any of the findings of this study change when the visual fidelity of VR avatars reaches that of video.

ACKNOWLEDGMENTS

This work would not have been possible without the support of Ronald Mallet, Reena Philip and Elif Albuz. It was informed by useful discussions with Ronit Kassis. Sohail Shafii, Carsten Stoll, Michael Ranieri and David Altman contributed to software development and technical support. Jeremy Schichtel provided support with technology and experiment execution. Aaron Ferguson, Victor Knai and Thierry DiDonna assisted with art assets. Many others contributed to the underlying technologies.

REFERENCES

[1] Jeremy N Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior* 2, 1 (2021).

[2] Jeremy N Bailenson, Andrew C Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. 2005. Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. *Human communication research* 31, 4 (2005), 511–537.

[3] Jeremy N Bailenson, Nick Yee, Jim Blascovich, Andrew C Beall, Nicole Lundblad, and Michael Jin. 2008. The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *The Journal of the Learning Sciences* 17, 1 (2008), 102–141.

[4] H Harlyn Baker, Nina Bhatti, Donald Tanguay, Irwin Sobel, Dan Gelb, Michael E Goss, W Bruce Culbertson, and Thomas Malzbender. 2005. Understanding performance in coliseum, an immersive videoconferencing system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 1, 2 (2005), 190–210.

[5] Douglas Bates et al. 2005. Fitting linear mixed models in R. *R news* 5, 1 (2005), 27–30.

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[7] Geoffrey W Beattie. 1979. Contextual constraints on the floor-apportionment function of speaker-gaze in dyadic conversations. *British Journal of Social & Clinical Psychology* (1979).

[8] Geoffrey W Beattie. 1981. The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. (1981).

[9] Gary Bente, Sabine Rüggenberg, Nicole C Krämer, and Felix Eschenburg. 2008. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research* 34, 2 (2008), 287–318.

[10] 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.

[11] Leanne S Bohannon. 2010. *Effects of video-conferencing on gaze behavior and communication.* Ph.D. Dissertation. Rochester Institute of Technology.

[12] Judee K Burgoon, Deborah A Coker, and Ray A Coker. 1986. Communicative effects of gaze behavior: A test of two contrasting explanations. *Human Communication Research* 12, 4 (1986), 495–524.

[13] Milton Chen. 2002. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 49–56.

[14] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[15] Richard L Daft and Robert H Lengel. 1986. Organizational information requirements, media richness and structural design. *Management science* 32, 5 (1986), 554–571.

[16] Trevor J Dodds, Betty J Mohler, and Heinrich H Bülthoff. 2011. Talk to the virtual hands: Self-animated avatars improve communication in head-mounted display virtual environments. *PloS one* 6, 10 (2011), e25759.

[17] Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'malley, Steve Langton, Simon Garrod, and Vicki Bruce. 1997. Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of experimental psychology: applied* 3, 2 (1997), 105.

[18] Wei Dong and Wai-Tat Fu. 2012. One piece at a time: why video-based communication is better for negotiation and conflict resolution. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 167–176.

[19] Henry Fuchs, Andrei State, and Jean-Charles Bazin. 2014. Immersive 3d telepresence. *Computer* 47, 7 (2014), 46–52.

[20] Susan R Fussell and Leslie D Setlock. 2014. Computer-mediated communication. (2014).

[21] Jonathan Gratch, David DeVault, Gale M Lucas, and Stacy Marsella. 2015. Negotiation as a challenge problem for virtual humans. In *International Conference on Intelligent Virtual Agents*. Springer, 201–215.

[22] Simon NB Gunkel, Hans M Stokking, Martin J Prins, Nanda van der Stap, Frank B ter Haar, and Omar A Niamut. 2018. Virtual Reality Conferencing: Multi-user immersive VR experiences on the web. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 498–501.

[23] J Richard Hackman. 1968. Effects of task characteristics on group products. *Journal of Experimental Social Psychology* 4, 2 (1968), 162–187.

[24] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. 2020. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 87–1.

[25] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

[26] Jonathon D Hart, Thammathip Piumsomboon, Louise Lawrence, Gun A Lee, Ross T Smith, and Mark Billinghurst. 2018. Emotion sharing and augmentation in cooperative virtual reality games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 453–460.

[27] 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*. 413–422.

[28] Paul Heidicker, Eike Langbehn, and Frank Steinicke. 2017. Influence of avatar appearance on presence in social VR. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 233–234.

[29] Kori Inkpen, Rajesh Hegde, Mary Czerwinski, and Zhengyou Zhang. 2010. Exploring spatialized audio & video for distributed conversations. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 95–98.

[30] 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.

[31] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. 2017. Effects of avatar and background types on users' co-presence and trust for mixed reality-based teleconference systems. In *In Proceedings the 30th Conference on Computer Animation and Social Agents*. 27–36.

[32] Tuomas Kantonen, Charles Woodward, and Neil Katz. 2010. Mixed reality in virtual world teleconferencing. In *2010 IEEE Virtual Reality Conference (VR)*. IEEE, 179–182.

[33] Sara Kiesler. 1986. *The hidden messages in computer networks*. Harvard Business Review Case Services.

[34] Chris L Kleinke. 1986. Gaze and eye contact: a research review. *Psychological bulletin* 100, 1 (1986), 78.

[35] Robert E Kraut. 2003. Applying social psychological theory to the problems of group work. *HCI models, theories and frameworks: Toward a multidisciplinary science* (2003), 325–356.

[36] Robert E Kraut, Susan R Fussell, Susan E Brennan, and Jane Siegel. 2002. Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. *Distributed work* (2002), 137–162.

[37] J. C. Lafferty, Eady, and J. Elmers. 1974. *The desert survival problem*. Plymouth, Michigan: Experimental Learning Methods.

[38] Thomas Maran, Marco Furtner, Simon Liegl, Theo Ravet-Brown, Lucas Haraped, and Pierre Sachse. 2020. Visual attention in real-world conversation: Gaze patterns are modulated by communication and group size. *Applied Psychology* (2020).

[39] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

[40] David McNeill. 2008. *Gesture and thought*. University of Chicago press.

[41] David T Nguyen and John Canny. 2007. Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1465–1474.

[42] David T Nguyen and John Canny. 2009. More than face-to-face: empathy effects of video framing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 423–432.

[43] 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.

[44] Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce. 1996. Comparison of face-to-face and video-mediated interaction. *Interacting with computers* 8, 2 (1996), 177–192.

[45] Charles A O'Reilly and Karlene H Roberts. 1976. Relationships among components of credibility and communication behaviors in work units. *Journal of Applied Psychology* 61, 1 (1976), 99.

[46] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 741–754.

[47] Ye Pan and Anthony Steed. 2016. A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise. *Frontiers in Robotics and AI* 3 (2016), 12.

[48] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1716–1725.

[49] Iasonas Kokkinos Rıza Alp Güler, Natalia Neverova. 2018. DensePose: Dense Human Pose Estimation In The Wild. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] Daniel Roth, Gary Bente, Peter Kullmann, David Mal, Chris Felix Purps, Kai Vogeley, and Marc Erich Latoschik. 2019. Technologies for social augmentations in user-embodied virtual reality. In *25th ACM Symposium on Virtual Reality Software and Technology*. 1–12.

[51] L Russell, S Henrik, L Jonathon, B Paul, and H Maxime. 2018. Estimated marginal means, aka least-squares means. *The American Statistician* 34 (2018), 216–221.

[52] Ralph Schroeder. 2011. Comparing avatar and video representations. In *Reinventing Ourselves: Contemporary Concepts of Identity in Virtual Worlds*. Springer, 235–251.

[53] Shayle R Searle, F Michael Speed, and George A Milliken. 1980. Population marginal means in the linear model: an alternative to least squares means. *The American Statistician* 34, 4 (1980), 216–221.

[54] Chris Segrin. 1993. The effects of nonverbal behavior on outcomes of compliance gaining attempts. *Communication Studies* 44, 3-4 (1993), 169–187.

[55] Leslie D Setlock, Pablo-Alejandro Quinones, and Susan R Fussell. 2007. Does culture interact with media richness? The effects of audio vs. video conferencing on Chinese and American dyads. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 13–13.

[56] Mel Slater, Amela Sadagic, Martin Usoh, and Ralph Schroeder. 2000. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators & Virtual Environments* 9, 1 (2000), 37–51.

[57] Harrison Jesse Smith and Michael Neff. 2018. Communication behavior in embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 289.

[58] William Steptoe, Anthony Steed, Aitor Rovira, and John Rae. 2010. Lie tracking: social presence, truth and deception in avatar-mediated telecommunication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1039–1048.

[59] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 197–200.

[60] Susan G Straus. 1996. Getting a clue: The effects of communication media and information distribution on participation and performance in computer-mediated and face-to-face groups. *Small group research* 27, 1 (1996), 115–142.

[61] Susan G Straus and Joseph E McGrath. 1994. Does the medium matter? The interaction of task type and technology on group performance and member reactions. *Journal of applied psychology* 79, 1 (1994), 87.

[62] W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S* (fourth ed.). Springer, New York. https://www.stats.ox.ac.uk/pub/MASS4/ ISBN 0-387-95457-0.

[63] Peter H Waxer. 1977. Nonverbal cues for anxiety: An examination of emotional leakage. *Journal of abnormal psychology* 86, 3 (1977), 306.

[64] Steve Whittaker. 2003. Theories and Methods in Mediated Communication: Steve Whittaker. In *Handbook of discourse processes*. Routledge, 246–289.

[65] Nelson Wong and Carl Gutwin. 2014. Support for deictic pointing in CVEs: still fragmented after all these years'. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1377–1387.

[66] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[67] Boram Yoon, Hyung il Kim, Gun A. Lee, Mark Billinghurst, and Woontack Woo. 2019. The Effect of Avatar Appearance on Social Presence in an Augmented Reality Remote Collaboration. In *IEEE VR*. ACM.

Fig. 9. Visualization of gaze tracking rays.

APPENDIX

Processing Gaze Data

Eye tracking in VR was performed with additional cameras mounted in the headset. For VC, gaze was recorded with a Tobii eye-tracker mounted on a stand in front of participants.

VR. Each participant's avatar in the VR condition had colliders assigned to the different avatar body parts. For the FloorPlan and PartyPlanning tasks, the "task objects" (i.e. virtual scoreboard and floor plan on table) were also assigned colliders.

Using logged pose data, we reconstructed each recorded frame from each experiment session. This pose reconstruction sets the eyeball orientation. A sphere-cast is performed from each eye, which sends a sphere along a ray in the view direction and logs all intersecting collider objects and their distance. To confirm integrity of the technique, visual inspections of a sampling of data were performed (e.g. Figure 9). During this process, some instances were noticed of gaze rays barely missing the body colliders, especially around the neck area, even though it looked like participants were focusing on those body parts. To reduce this error, we enlarged the colliders between 10 and 20%.

VC. Tobii screen captures with associated eye-tracking data were exported for automatic labeling using an external tool developed for this purpose. As additional input for this tool, we first manually defined rectangular regions of interest (ROIs) for each screen capture. These ROIs describe in screen-capture space where each participant's video streams were, and where task-related information was shown. Next, our tool automatically annotated the participant ROIs with body part segmentation masks using the detectron2/densepose network [49, 66]. This mask assigns a label to each pixel with either *background* or with the body part inferred by the network. Body part categories matched those used in VR. [1]. Next, the Tobii eye-tracking coordinates in screen-capture space are associated with each frame and their intersection with body parts masks are determined. To be consistent with the sphere cast technique used in VR labeling, a circle of comparable size was placed at the gaze location and the label was assigned based on a majority vote of the pixels in this circle.

[1]Given the amount of data to be processed, and this process being computationally expensive, the body part segmentation masks were only updated every fifth frame.

Post-processing: Every gaze sample was labeled with 1) who it was from 2) whether the gaze was at the `Body`, a `Task` object or `Elsewhere`, and 3) body gaze was broken down to `Head, Torso, Hands and Lower` categories. For VR, there are two samples for each pose frame (one for each eye) and the one with smallest depth was selected.

Equalizing Gaze Data Across Media: Because VR gaze is tracked in the 3D scene, data is available wherever the person looks. For VC, however, gaze is only tracked when they look at the screens. No data for VC gaze means either that the participant is looking off-screen or the gaze tracker failed. To understand which event is likely, a sampling of 120 gaps of varied duration with no tracking data were randomly selected and manually coded based on a video review with a 1 if the participant was gazing off-screen and 0 if the participant was looking at the screen, but tracking had failed. Gaps $\leq 0.5s$ generally correspond to blinks and short term tracking failures, so were filled in with the surrounding tracking labels. Gaps of longer than 100s never consisted of participants looking off-screen, so these were considered failed tracking and discarded. Segments in between the two extremes were a mix of participants looking off-screen or tracking failures (often caused by participants shifting in their seats). Longer segments were more likely to be failed tracking, but there were cases of both off screen gaze and failed tracking throughout the range. To account for this, a regression line was fit to the 1 or 0 labels and used to indicate the probability that sample was valid off-screen data (below the line) or failed tracking (above). The equation of the line was $y = 0.8605 + -0.007278x$ where x represents the duration of the segment. Gaze marked off-screen was included in the "Elsewhere" category. While an individual case may be wrong, this provides a statistically reasonable categorization of gaze.

Unlike VC, VR gaze tracking can detect when a person is looking at their own body. A common VR gaze target was the users own hand (9.3% of gaze), which sometimes occurred when they were using a virtual task interface. To further make VR and VC more comparable, we opted to ignore self-hand gaze collisions and instead take the next collision behind the hand. This was a Task target 30-40% of the time, and "Elsewhere" the rest.

Nonverbal Annotation Process

After the study sessions, remote annotators annotated the groups' nonverbal behavior during the floor plan negotiation task using an audio and video feed that displayed all participants. Annotators labeled the type of every gesture performed, with types shown in Table 3. In addition, they could indicate if the gesture contained unique information with the options in Table 4. Mouse movement in VC was included in the annotation as a substitute for manual gesture (e.g. deixis and spatial gestures were often performed with the mouse).

All annotations for the project were completed by a team of four annotators who saw the videos in random order. Every video was independently annotated by two different annotators. If there were mismatches between their ratings, they were resolved by a third annotator. The research team performed quality checks throughout the annotation process.

TASK INSTRUCTIONS

Estimation

In this task, you will be asked a series of questions where you will need to estimate some real world statistics, for example, what the median age is in the U.S. Your bonus of up to $11 for this task will depend on how many questions you answer correctly. This task will last 15 minutes. After each question, please discuss your answer with the group. When you've agreed, report your answer

Gesture Type	Description
Backchannel	Acknowledgments of interlocutor, including head nods and manual gestures.
Beat	Small movements of the hand in rhythm with the spoken prosody.
Deictic	Pointing gesture.
Emblem	Well defined gestures used as a substitute for words, such as a thumbs up.
Glitch (in VR)	Tracking failure generating an unnatural movement.
Regulator	A gesture that passes the conversational floor to the person who should speak next.
Representation	Metaphoric and iconic hand movements, illustrative of an idea (but not fitting in "Spatial or Distance").
Self-adaptor	Self-manipulations not designed to communicate, such as nose scratches.
Spatial or Distance	Gestures conveying more complex spatial or distance information, such as a path through the apartment.

Table 3. Annotators applied the most appropriate label to each observed gesture.

Novel Content	Description
Reference Pronoun	Use of pronouns that must be disambiguated by the gesture, such as saying 'this' while pointing.
Unique Information	Information other than a reference pronoun that is only provided through the gesture
Default	Applied if no other option was selected.

Table 4. Annotators could apply additional metadata about each gesture.

by saying "Final Answer" and let me know what the answer is. You only have one chance to give an answer for each question. If you discuss a question for more than 3 minutes, I will give you a 15 second warning and then move on to the next question. I will not be answering any questions during the task, so please ask questions now if you are unclear about anything. Any questions?

Ok. Let's do a quick check for understanding.

- How will your bonus be determined for this task?
- Does everyone need to agree with the answer?

Bribery

In this scenario, you are all members of a personnel discipline committee at a major company. An employee we'll call "Salesperson" works in sales. Salesperson's boss we'll call "Boss" and is head of the sales department. Salesperson is one of the top performers in the company, ranking in the top 5% of sales. Last year Boss found out that Salesperson had accepted an expensive trip from one of the company's clients, against company policy. Salesperson apologized, promised to never do it again and pleaded to avoid punishment. Company policy requires Boss to report this offence to his or her boss. Afraid of losing a top producer, Boss did not report Salesperson and let Salesperson off with a warning. A whistle blower has reported the issue and it has been turned over to your committee. In your deliberations, please consider three groups with potentially conflicting interests:

- The sales team, that wants to keep their star.
- The CEO, who is worried about insubordination and the negative impact an ethics violation would have if this reached the press.

- The overall company culture.

You need to decide on four issues:

- The salesperson's punishment.
- Whether the salesperson can keep their job
- The boss's punishment
- Whether the boss can keep their job

Examples of punishment include:

- No penalty
- Stiff warning
- Stiff warning and probation
- Fine equal to the cost of the trip
- Fine double the cost of the trip
- Loss of job

You must collectively decide on the resolution for each issue and you must agree on a justification for each decision. Your bonus of up to $11 will depend on the quality of the plan you develop as a group, as well as how much you contributed to the discussion. Your group must come to an agreement in order to receive a bonus. This task will last 15 minutes and I will give you a 2 minute warning towards the end. Do you have any questions?

Ok. Let's do a quick check for understanding.

- What did Salesperson do wrong?
- What did Boss do wrong?
- What three groups with potentially conflicting interests should you consider?
- What are some example punishments you might consider?
- Are you allowed to consider other punishments?
- What are the four issues you must resolve?

Party Planning

You are each senior managers at a major company. One of you is Head of Finance, one of you is Head of Security, and one of you is Head of Social Events. You have been given the task of planning the end of year party for 500 people and you need to reach a collective decision on each of four issues:

- How many knife jugglers should you hire to perform at the party? Your options are 1, 2, 3 or 4.
- How many security guards should you hire for the party? Your options are 5, 10, 15 or 20.
- How much should you charge employees to bring a guest? Your options are $0, $10, $20 or $30.
- What time should the party end? Your options are 8pm, 9pm, 10pm or 11pm.

Each of you have preferences for each of these items, based on the work unit you represent. We will give you your preferences separately. Your job on this task is to negotiate with each other to reach a final decision on each of these items. Your bonus for this task will be up to $11 and will depend on how many of the total points you receive. Once the task begins, you will have 15 minutes to negotiate and I will give you a three minute warning towards the end. If you do not reach an agreement, no one will receive any bonus. Do you have any questions?

Ok. Let's do a quick check for understanding.

- What are the four topics you are negotiating on?

- What are your options for the amount of knife jugglers you can have at the party?
- What are your options for the amount of security guards to hire at the party?
- What are your options for the amount for the ticket cost for guests?
- What are your options for what time the party should end?
- What does your bonus depend on?
- What is your bonus if you do not reach agreement?

Floorplan

The three of you have agreed to be roommates. The floorplan shows the apartment you have agreed to rent. Study it closely to understand the different features of the apartment. There are three bedrooms, two with large windows, two bathrooms, a living room and a kitchen. Two bedrooms have large windows, one of the two also has a view. You need to decide who gets each bedroom and which room will be a living room. In addition, there is a closet in the common space that can be assigned to one of the roommates. The total rent is $3,000. You need to decide how to fairly split the rent and best share the space.

As in most situations, each of you value different things. We'll give you a reward table that summarizes what's important to you. Do not share or talk about this table. You should negotiate with each other to decide on the room allocation and what portion of the rent each of you should pay. Try to make convincing arguments. Your potential bonus of up to $11 for this task will depend on how well you meet your goals.

You each have been given your personal reward tables. You should not directly reveal the numbers in this table. You have two minutes to study the information and come up with the best arguments you can for why you should get the features you want. You are not allowed to look at these papers during the task. If you forget the information, you will be able to click on a button to reveal it while negotiating. Any questions?

Ok. Let's do a quick check for understanding.

- What is the total rent of the apartment?
- What are the four issues you be negotiating?
- How is your bonus determined for this task?
- What happens if you do not come to an agreement by the time is up?

Received April 2021 ; accepted July 2021