Journal on Multimodal User Interfaces
https://doi.org/10.1007/s12193-020-00330-2

ORIGINAL PAPER

Sharing gaze rays for visual target identification tasks in collaborative augmented reality

Austin Erickson[1] · Nahal Norouzi[1] · Kangsoo Kim[1] · Ryan Schubert[1] · Jonathan Jules[1] · Joseph J. LaViola Jr.[1] · Gerd Bruder[1] · Gregory F. Welch[1]

Received: 8 January 2020 / Accepted: 6 June 2020

Abstract

Augmented reality (AR) technologies provide a shared platform for users to collaborate in a physical context involving both real and virtual content. To enhance the quality of interaction between AR users, researchers have proposed augmenting users' interpersonal space with embodied cues such as their gaze direction. While beneficial in achieving improved interpersonal spatial communication, such *shared gaze environments* suffer from multiple types of errors related to eye tracking and networking, that can reduce objective performance and subjective experience. In this paper, we present a human-subjects study to understand the impact of *accuracy*, *precision*, *latency*, and *dropout* based errors on users' performance when using shared gaze cues to identify a target among a crowd of people. We simulated varying amounts of errors and the target distances and measured participants' objective performance through their response time and error rate, and their subjective experience and cognitive load through questionnaires. We found significant differences suggesting that the simulated error levels had stronger effects on participants' performance than target distance with accuracy and latency having a high impact on participants' error rate. We also observed that participants assessed their own performance as lower than it objectively was. We discuss implications for practical shared gaze applications and we present a multi-user prototype system.

Keywords Shared gaze · Eye tracking · Eye tracking errors · Collaborative augmented reality · Target identification

1 Introduction

Over the last several years, great strides have been made to improve sensor and display technologies in the fields of augmented reality (AR) and virtual reality (VR) [19]. These advances, such as with respect to head-mounted displays (HMDs) and eye trackers, have provided new opportunities for applications in fields such as training, simulation, therapy, and medicine. For many of these, collaboration between multiple users is an important aspect of the experience. In real life, people use both verbal and nonverbal cues to communicate information to the person they are interacting with.

In order to understand and improve collaborative experiences using AR/VR technologies, researchers have evaluated the impact of different embodied and behavioral cues on users' efforts and performance [21,32,34]. Researchers have also investigated how certain embodied cues could be augmented to improve their efficiency in interpersonal communication. A prime example of such augmented cues are *shared gaze environments* [32]. Eye gaze is an important cue for spatial interaction and collaboration among humans as it

Austin Erickson and Nahal Norouzi have contributed equally to this research.

✉ Austin Erickson
ericksona@knights.ucf.edu

Nahal Norouzi
nahal.norouzi@knights.ucf.edu

Kangsoo Kim
kangsoo.kim@ucf.edu

Ryan Schubert
rschuber@ist.ucf.edu

Jonathan Jules
julesj@knights.ucf.edu

Joseph J. LaViola Jr.
jjl@eecs.ucf.edu

Gerd Bruder
bruder@ucf.edu

Gregory F. Welch
welch@ucf.edu

[1] The University of Central Florida, 3100 Technology Parkway, Orlando, FL 32826-3281, USA

Published online: 09 July 2020

lets us know what another person is looking at, which is often in line with their focus of attention [23]. Gaze cues further inform us about a person's mental processes, eye contact, and gaze avoidance. *Joint gaze* occurs when multiple people are looking at the same object when it is the focus of discussion [20].

Advances in AR requires a better understanding of its interaction space [41], such as AR technologies' potential to augment such eye gaze cues in interpersonal communication, e.g., by providing a *gaze ray* from the user's head to the object in the environment that they are looking at. Different methods have been proposed to share gaze rays, cones, or focus points between users [21,34,42]. Some of these approaches require information about the geometry of the physical environment, while the most basic and generalizable approaches are mainly based on rays that indicate the direction of gaze but do not terminate at any object.

All shared gaze environments have to track users' gaze direction with an eye tracker, e.g., integrated into a HMD, and they have to transmit that information for subsequent rendering in their AR view, e.g., via a wireless network. The quality of shared information and the errors involved in this process are highly important to quantify and understand for practitioners in shared spaces [8,35,42].

In this journal article, we resume and extend work originally published in an impactful conference publication by Norouzi et al. [28], presenting a human-subject study aimed at understanding the importance and influence of four types of errors in AR shared gaze environments on users' performance and perception. We focused on *accuracy* and *precision* related to the eye tracker as well as *latency* and *dropout* related to the network. We created a scenario where a participant had to collaborate with a simulated partner to identify a target among a crowd of humans. Inspired by related work [21,34,42], we augmented the participant's view with a gaze ray that indicated which person their simulated partner was looking at. We measured participants' performance using response time and error rate. Our findings suggest that participants performed surprisingly well, better than they believed themselves when we asked them to judge their performance, even when the simulated error levels were high, with accuracy and latency having the highest impact on participants' error rates. We discuss the results for practical shared gaze applications and present a prototype of an AR multi-user shared gaze system.

With this work, we aim to contribute to the research community by providing answers to the research questions below:

- **RQ1** How do different types of errors affect users' collaborative performance in AR shared gaze environments?
- **RQ2** What are the thresholds for the amounts of error introduced without affecting users' performance?
- **RQ3** How different are users' subjective and objective assessment of the experience and what is their subjective perception of an acceptable amount of error for the task at hand?

This paper is structured as follows. Section 2 discusses related work. Section 3 describes the experiment. Section 4 describes our results and Sect. 5 discusses our findings. We further present and discuss a prototype AR shared gaze system in Sect. 6. Section 7 concludes the paper and discusses future opportunities for research.

2 Related work

In this section, we present related work on collaborative shared spaces, shared gaze cues, and errors impacting user experience and performance.

2.1 Sharing gaze in AR/VR

In an early work by Kiyokawa et al., a mixed-space collaborative platform was introduced that included an awareness enhancing technique to improve the quality of collaboration between the two users [21]. This was achieved by visualizing a gaze ray initialized from between the user's eyes, finding that participants rated their task to be easier with the gaze ray when one participant had to guide the other in finding stationary targets. Similarly, Bauer et al. investigated the effects of a "reality augmenting telepointer" used for expert-worker scenarios for mobile workers. Using their system, the expert's pointer was displayed to the user [2]. They found that with the inclusion of the pointer, similar speech behaviors were observed as in face-to-face conversations. Piumsomboon et al. introduced a system called COVAR that could facilitate collaboration between AR and VR users and is able to share their head frustum related to their field of view, head-ray, gaze-ray, and hand gestures of users with each other to improve the collaborative experience [34]. Their results emphasize the positive impact of these cues on aspects such as performance and subjective preference. Brennan et al. compared different combinations of gaze and voice cues where remote users took part in a search task finding that sharing gaze information alone resulted in faster search times than gaze and voice conditions [3]. In a helper-worker scenario, Gupta et al. looked at effects of sharing the worker's gaze with the helper and the helper's pointer with the worker. Their results indicate positive impacts of having both cues on performance and quality of the experience [13]. With such findings emphasizing the benefits of shared gaze, Zhang et al. investigated the impact of the methods used to visualize the gaze (i.e., highlight, spotlight, trajectory, and cursor) on users' performance and cognitive load [42]. Their find-

ings suggest that users perceived the highlight and spotlight modes as less distracting compared to other modes such as cursor and trajectory. Although past research have had valuable contributions in understanding the influence of sharing gaze in AR/VR for collaborative purposes, to our knowledge, the influence of different types of errors inherent to shared gaze experiences have not been studied. Knowing these influences exist in single user experiences, a better understanding is required for shared experiences.

2.2 Gaze tracking performance

As the quality of the data reported by eye trackers is dependent on various factors, understanding the causes and effects of these issues on user performance is important for practical applications. Holmqvist et al. discussed basic examples of different types of errors in eye tracking environments that are caused by limitations of current-state eye trackers (e.g., accuracy and precision), and how these can cause misinterpretations in different measures such as dwell time [16]. They discussed factors that can affect the quality of the data such as the experimental task and eye tracking algorithms. Other researchers also provided explanations for sources of eye tracking errors such as variations in pupil size [7], eye color as well as calibration instructions and methods [29].

Some researchers proposed methods to compensate for these errors. For instance, Cerrolaza et al. proposed calibration techniques to compensate for the impact of user movement on the devices' accuracy [5]. For less expensive commercial off-the-shelf (COTS) eye trackers, Ooms et al. proposed steps to improve their data quality [30]. Hornoff et al. proposed using the disparity between the true position of implicit fixation points and the reported value from the eye tracker as a way to measure the robustness of the reported data [17]. Barz et al. proposed a computational approach that would model and predict gaze estimation errors in real time and could be used in applications to identify high error regions during user interaction and modify the elements such as increasing the size of the objects [1]. While varying factors such as lighting and eye tracker type, Feit et al. found large differences in eye tracking data quality and proposed new design choices such as target placement and size adjustments to compensate for these variabilities [10].

With most of past work's focus on assessing the gaze tracking errors, identifying contributing factors and solutions reducing their impact, further investigation of these errors in more dynamic shared AR/VR setups can be beneficial.

2.3 Network sharing performance

Network performance in terms of transmission latency and dropout is an important factor that shapes user performance in shared AR/VR experiences. A large body of literature showed that latency has a negative effect on user performance, but most of that research focused on tracker or rendering latency between a person's physical movements and the computer-generated feedback [8,18,22,24,31,35]. In contrast, effects of network latency and dropout in shared AR/VR environments have less immediate cause-effect relationships.

We are not aware of previous research investigating latency and dropout in shared gaze environments, but related work in AR/VR and general communication focused on aspects of collaborative environments [12,36]. Recently, Toothman and Neff investigated different network errors in an embodied multi-user VR setup, including latency and dropout and their effects on social presence [38]. Their results showed practical thresholds such as that a latency of 300 ms and dropout with freezing frames for 100–350 ms for 67% of the time had a negative impact on users' experience and performance. dropout (or frame dropping) has been further researched by Pavlovych et al., who identified a threshold of 10% frame drops, after which it had a negative effect on participants' tracking task performance on a computer [31]. Geelhoed et al. showed that the conversation flow in a telepresence system was reduced by added network latency, recommending a limit of 100–600 ms for round trip time latency, but they also found that basic conversations without time sensitive tasks were not that affected by latency and could go up to 2000 ms of latency [12]. Other research showed that network latency might further lead to misinterpretations of users' dispositions during interactions [36].

Further investigation of different error types for collaborative purposes can provide a better understanding of their implications on users' collaborative performance and experience.

3 Experiment

In this section, we present the experiment that we conducted to assess the impact of different types of errors that are inherent to collaborative shared gaze environments in AR.

3.1 Participants

We recruited 21 participants (7 female, 14 male, age 19–36, average 23.28) from the graduate and undergraduate population of our university. The protocol for our experiment was approved by the institutional review board (IRB) of our university. All participants indicated normal hearing and normal/corrected vision. Before the experiment, we asked our participants to use a 7-point scale (1 = novice/unfamiliar, 7 = expert/familiar) to rate their familiarity with AR (average 4.7), VR (average 5.19), virtual humans (average 3.9), and overall computer expertise (average 5.52).

3.2 Material and task description

We conducted the experiment in an open 4.6 m × 10.4 m space in our laboratory. We used two computers with Intel Xeon 2.4 GHz processors comprising 16 cores, 32 GB of main memory and two Nvidia Geforce GTX 980 Ti graphics cards for the stimulus control and for participants to answer questionnaires. We used the Unity graphics engine version 2018.2.11f1 for rendering, and a Microsoft HoloLens for the presentation of the visual stimulus.

We tasked the participants with working with a virtual human partner who took the form of a police officer. The participant and their virtual partner are tasked with identifying a threat in the crowd of moving virtual humans, and the virtual officer communicates the presence of this threat nonverbally by sharing his gaze information with the participant. It is the participants' goal to utilize this gaze information in order to identify who the potential threat is out of all of the virtual humans in the crowd.

3.2.1 Shared gaze stimuli

To provide repeatable controlled shared gaze stimuli in a manner similar to what has been previously done by Murray et al. [27], we decided to use a simulated virtual human partner in this study. We placed a 3D virtual human character (see Fig. 1) at a distance of one meter on the left side to the participants, which was visible to them on the HoloLens. During the experiment, the simulated partner stood with an idle animation, facing in the same forward direction as the participants. A 20-meter gaze ray was presented in AR that originated in the partner's eye location and went forward into the environment. The gaze ray was programmed to be rendered on top of the real or virtual entities in the environment to be more in line with practical shared gaze setups and to not give away or misrepresent a target through depth cues resulting from an intersection with a target at any moment. We would like to point out that such gaze rays are mainly used when one does *not* have access to a high-precision real-time reconstruction of the geometry of a physical environment, as discussed in Sect. 1.

We tested different gaze simulation algorithms but noticed that these were largely not able to create realistic gaze behavior in AR. In order to create a natural gaze behavior for the simulated partner, we recorded the eye behavior of one of the experimenters looking at a stationary target located at a distance of one meter away. We used the Pupil Labs[1] software to capture the recording and assess its accuracy (0.55 deg) and precision (0.08 deg). This recorded data was then analyzed to find the average gaze position observed (with the accuracy

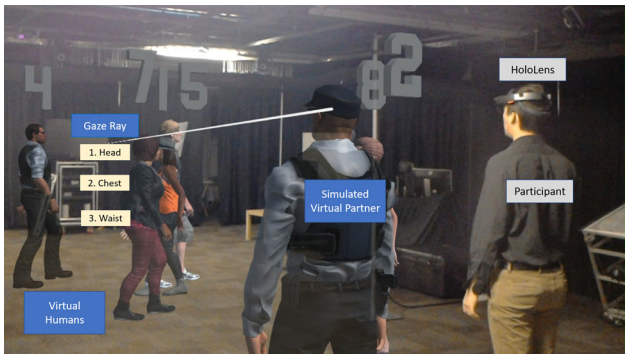


Fig. 1 Annotated screenshot showing a participant wearing a HoloLens with the simulated virtual partner on their left side, looking at a target in a crowd of virtual humans. The virtual target humans are differentiated by the floating numbers above their head

error), and was then normalized around this position to yield data with no accuracy error.

This recorded gaze data was played back in Unity and oriented to simulate saccades and smooth pursuit movements to follow target points on the moving target humans in the environment. Each of them had three points of interest, one on their head, one on their chest, and one near their waist (see Fig. 1).

The script that controlled the gaze behavior would target one of these points at random every 750 ms, with a fifty-percent probability of choosing the head as the target and a twenty five percent chance of choosing either of the other two points. This behavior made the virtual partner's gaze seem as though it was identifying the target human by recognizing their face, while scanning the target for concealed weapons. The simulated gaze followed the targets when they were moving.

3.2.2 Gaze target crowd

Our setup consisted of nine simulated virtual human targets shown in Figs. 1 and 2. The 3D models and animations were acquired through the Unity Asset Store[2] or Mixamo.[3] The virtual humans (4 female, 5 male) were placed 0.7 m apart from each other in depth. Walking animations were added to each model so they could pace back and forth between two predefined points on the left and right sides from the participants with a total distance of 6 m. Each virtual human was presented with a floating number over their head to make it easier for the participants when reporting the gaze target. The walking speed for the virtual humans was chosen from the range 0.8 ± 0.2 m/s, which is close to average human walking speed [11]. Target virtual humans could be chosen

[1] https://pupil-labs.com/.

[2] https://assetstore.unity.com/.

[3] https://www.mixamo.com/.

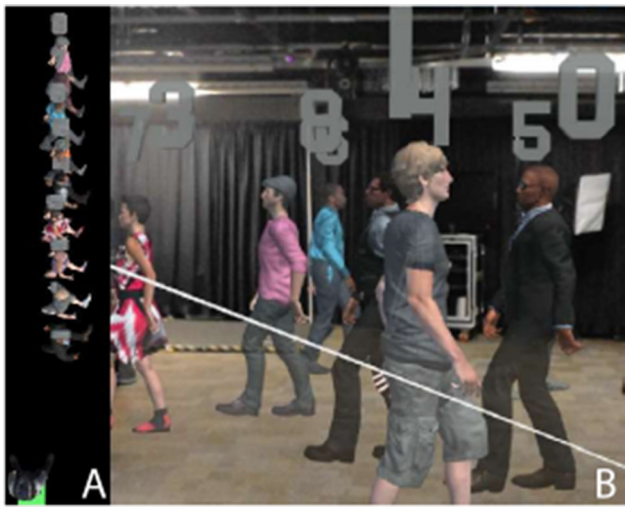


Fig. 2 **a** Shows a top-down view of the starting configuration of the virtual humans from within the Unity editor. The virtual humans appear in a randomized order with randomized speeds for each trial. **b** Shows a view of the virtual humans from the participants' perspective after the crowd has started moving. In this case, the virtual partner is observing the waist of virtual human 1

from one of three different distances (see below) which corresponded to two different social proxemics categories as laid out by Hall, where the closest potential target fell into social distance while both other distances fell into public distance [14]. The order of the virtual humans in depth and their walking speed were randomized prior to commencing the first trial as well as between trials in the study (Fig. 2).

3.2.3 Gaze error implementation

As discussed in Sect. 1, we considered four different types of error that are common to eye trackers and shared gaze AR experiences. Figure 3 illustrates each error type in comparison to a no-error example. Below, we describe each error, its possible source and how it was implemented.

Accuracy: Persistent angular offset between the true eye gaze direction and the direction of the drawn gaze ray. To implement this error, the gaze ray for no accuracy error was calculated based on the simulated virtual partner's gaze direction and the recorded gaze data. This ray was then rotated towards the rightward horizontal axis of the target by a variable number between 0 and 5 degrees at 1 degree intervals to achieve an accuracy offset along the horizontal axis. It is important to note that while a physical eye tracking system would introduce errors in both the x and y directions, we opted to simulate the most extreme type of accuracy error that could occur for the study scenario, which was a horizontal shift away from the target. Due to the nature of our study scenario, a vertical offset would still appear on the target's body or slightly above the target's head, which would make the target easier to identify than if the shift had occurred in the horizontal direction alone. Additionally, a combination of these two directions would only limit the horizontal offset away from the target, and would also result in a target that is easier to identify than if the shift had occurred in solely the horizontal direction.

Precision: Dynamic angular differences between an eye tracker's reported eye gaze direction and the true direction to the gaze target. To implement this error, we calculated the gaze ray position based on the simulated virtual partner's gaze direction, then offset this position by an amount based on the recorded gaze data which was multiplied by a variable scale factor between 0 and 2.5 degrees at 0.5 degree intervals. This calculation would yield a gaze behavior that was centered on the target with increased variations around the target point as the scale factor increased.

Latency: End-to-end delay in the presentation of the gaze ray from the simulated partner's eyes. Here, our focus is on the latency introduced by the complex setup of a collaborative AR shared gaze system, which includes latency from the eye tracker, a wireless network, a rendering system, and a display. To realize this error, we computed the position of the target virtual human at a simulated temporal offset up to 1000 ms into the past. This past position was then set as the target for the simulated gaze ray. This was achieved by creating a dictionary that paired vector positions with timestamps for each virtual human in the scene. This dictionary could be searched to find a virtual human's past position based on the time difference between the current time and the amount of latency in milliseconds that was simulated.

Dropout: Here we define dropout as the probability of dropped or lost frames due to networking or eye tracking issues (e.g., eye tracker not being able to detect the pupil). To implement this error, for every gaze ray that was rendered we measured the chance of the next ray being dropped based on predefined values that are introduced in Sect. 3.3.1. As an example, if the predefined value was set to 90%, then there was a 90% chance that the next frame was dropped resulting in an inconsistent gaze ray.

To choose ecologically valid error ranges, we looked at the literature described in Sect. 2 and included the nominal performance reported by manufacturers for commercial head-worn eye trackers such as from Tobii[4] and Pupil Labs. To make sure that participants were able to perceive at least the maxima of all types of errors, we chose the maxima of our error ranges as slightly larger than the range of values reported in the mentioned sources.

[4] https://www.tobii.com/.

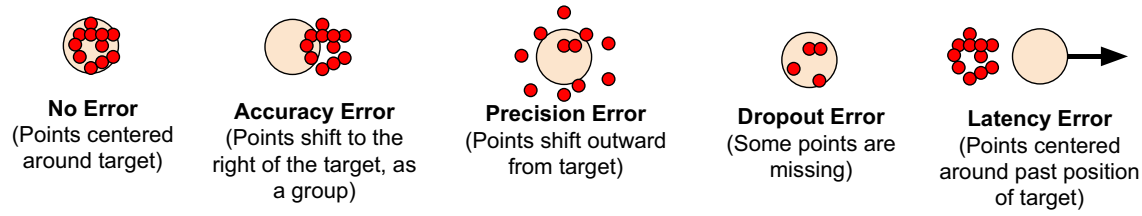


Fig. 3 Illustration of the different error types simulated in the shared gaze interface in comparison to a no-error example

3.3 Methods

3.3.1 Study design

We chose a $4 \times 6 \times 3$ within-subjects design for our experiment. This choice was made to account for the impact of individual differences on task performance. Our independent variables were:

- **Error type and error level** (4×6 factors):
  - **Accuracy:** 6 levels of accuracy error were introduced to the gaze ray from 0 deg to 5 deg with increments of 1 deg.
  - **Precision:** 6 levels of precision error were introduced to the gaze ray from 0 deg to 2.5 deg with increments of 0.5 deg.
  - **Latency:** 6 levels of latency error were introduced to the gaze ray from 0 to 1000 ms with increments of 200 ms.
  - **Dropout:** 6 levels of dropout error were introduced to the gaze ray starting at 0% and going from 10 to 90% with increments of 20%.
- **Target distance** (3 factors):
  - **Close:** The target was pacing back and forth at a distance of 3 m.
  - **Medium:** The target was pacing back and forth at a distance of 5.1 m.
  - **Far:** The target was pacing back and forth at a distance of 7.2 m.

These combinations of independent variables resulted in a total of 72 trials. For the experiment, we randomized the participants' exposure to each error type resulting in four combinations, and within each error type, we randomized their exposure to the *error levels* for different *target distances*, resulting in 18 combinations.

3.3.2 Procedure

After giving their informed consent, participants were asked to take a seat and answer a pre-questionnaire to rate their familiarity with AR, VR, computers, and virtual humans and to assess their hearing and vision. After this, they were asked to read a document about their task in the experiment and then the experimenter reviewed the written document with them.

Participants were guided to stand on the white cross marked on the floor a meter apart from the simulated AR collaboration partner and don the HoloLens. They were informed that they could move their head freely but that they should remain standing on that spot. Participants were asked to use the information from the partner's gaze ray to identify a potential suspect among the gaze target crowd in front of them.

Two distinct beep sounds were used to mark the start and end of each trial and participants were told that they should say their answers out loud by naming the number on top of the identified target human in the crowd. Participants were informed that they have a maximum of 60 s to make a decision for each trial but that they have to answer as quickly and confidently as possible. Upon reaching the 55th second in each trial, the virtual background turned red to indicate that the end of the trial was close and participants were asked to indicate their best guess.

After going over the procedure, participants took part in five practice trials. We then asked them if they think they need more practice. If their answer was *yes*, they took part in five more practice trials. Among our participants, only one asked for the extra five practice trials and we checked the response times for the 72 main trials for that participant to ensure that the extra five practice trials had not given them an advantage over the others.

After each block of 18 trials (i.e., one error type with six error levels and three target distances) they were asked to doff the HoloLens and answer a post-questionnaire. For each participant, this process was performed four times in randomized order within and between the blocks. After finishing the last block, participants were asked to answer further questionnaires. Then, they took part in a short debriefing session to discuss their experience. They received monetary compensation ($15) for their participation in the study.

3.3.3 Measures

In this section, we present the objective and subjective measures used to assess participants' performance and perception of each error type.

Objective measures As participants were asked to identify the potential gaze target as quickly and confidently as they can, we used these criteria to separately assess their performance in the form of **response time** and **error rate**. We would like to point out that although different models for speed-accuracy trade-offs have been introduced in the human-computer interaction literature, we are not aware of any validated model that could be applied to our specific stimulus-response case, such that we had to treat these measures separately:

- **Response time:** We recorded the amount of time taken for each participant to indicate a gaze target for each trial.
- **Error rate:** We recorded if participants identified the correct gaze target for each trial.

When participants made selection errors during the study, a log file recorded information about the participants' selected virtual human and the correct virtual human. This log recorded the z-depth in meters for both of these virtual humans, and speed in meters per second for both of these virtual humans. These pairs of depth and speed values could be compared to reveal insights into the selection strategy that participants used to make their decisions during the task.

Subjective measures We measured the subjective perception of our participants about the various error types and levels, and how their performance and general experience was impacted. The subjective measures used were:

- **Subjective performance:** For each block associated with a certain *error type*, three questions were used to assess participants' confidence in their answers (7-point Likert scale), their subjective performance (numeric response), and their subjective judgment of what constitutes an acceptable amount of error for the task at hand (numeric response). Table 1 shows these questions.
- **Subjective experience:** To understand the impact of our independent variables on how participants experienced each error type, we included questions from the NASA TLX cognitive load (CL) questionnaire [15], System Usability Scale (SUS) questionnaire [4], and a question asking about the realism of the gaze behavior. A 7-point Likert scale was used for all the questions in this questionnaire. Table 2 shows the questions used for this questionnaire.
- **Trust in Technology:** To assess participants' overall trust in technology and the shared gaze interface, we included several questions from McKnight et al.'s Trust in Technology questionnaire [26]. A 7-point Likert scale was used for all items and they were adjusted to match our interface. This questionnaire, shown in Table 3, was presented to the participants after they completed all four blocks.

3.3.4 Hypotheses

Based on the literature (see Sect. 2) and a hypothesis-generating pilot study with five participants (different from our study population, who generally made conservative estimates on their own performance), we formalized the following hypotheses:

- **H1:** Participants' **response time** and **error rate** will increase as the *error levels* increase within each *error type*.
- **H2:** Participants' **response time** and **error rate** will increase as the *target distance* increases within each *error type*.
- **H3:** Based on the inherent nature of *accuracy* and *latency* errors that provide a constant spatial and temporal offset, compared to *precision* and *dropout* errors, participants will:
  - **a** Give lower SP1 scores for the former error types and indicate less confidence in their answers,
  - **b** Give higher scores for CL1, CL2, CL3 for the former error types and indicate higher cognitive load,
  - **c** Give a lower SUS1 score for the former error types, assessing them as more difficult to use.
- **H4:** Participants' subjective estimate of the percentage of correctly identified targets answered through SP2 will be lower than their actual performance.

4 Results

In this section we present the objective and subjective results for our experimental conditions. For the analysis of our results, we removed the data of one of our participants as it failed our sanity checks; we noticed that several responses were for targets that were located in completely opposite places compared to the actual target.

4.1 Objective measures

We analyzed the results for the objective performance measures with repeated-measures ANOVAs and Tukey multiple

Table 1 The subjective *Performance* questionnaire

ID	Question
SP1	How confident were you on the correctness of your choices in this section of the experiment?
SP2	What percentage of the targets do you think you identified correctly from 0 to a 100%?
SP3	What do you think is an acceptable error margin for the system, based on your assessment of your performance?

Table 2 The subjective *Experience* questionnaire

ID	Question
CL1	How mentally demanding was the task?
CL2	How hard did you have to work to accomplish your level of performance?
CL3	How insecure, discouraged, irritated, stressed, or annoyed were you?
SUS1	I thought the system was easy to use
SE1	I felt the gaze behavior of my partner was realistic

Table 3 The adjusted *Trust in Technology* questionnaire

ID	Question
TT1	The shared gaze system is a very reliable piece of software
TT2	The shared gaze system has the features required for my task
TT3	I am totally comfortable working with the shared gaze system
TT4	I believe that most technologies are effective at what they are designed to do
TT5	I usually trust a technology until it gives me a reason not to trust it

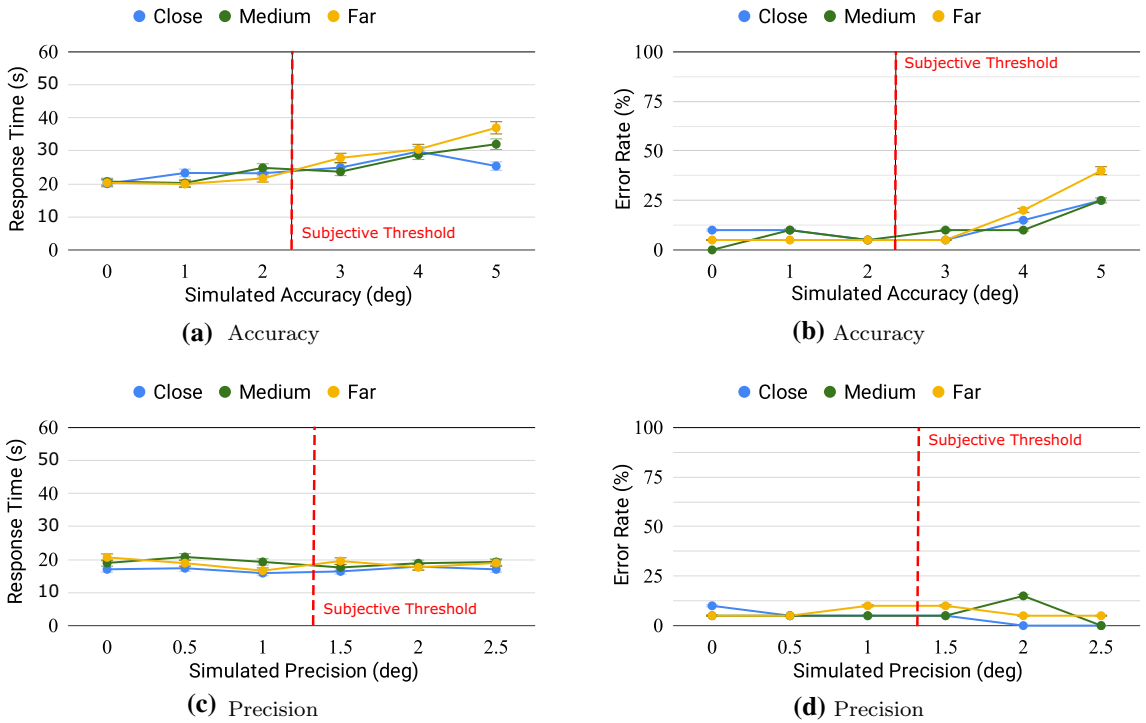


Fig. 4 Performance results related to eye tracking errors: The top row (**a**, **b**) shows results for *accuracy* and the bottom row (**c**, **d**) for *precision*. The left column (**a**, **c**) shows results for *response time* and the right column (**b**, **d**) for *error rate*

comparisons with Bonferroni correction at the 5% significance level. We confirmed the normality with Shapiro-Wilk tests at the 5% level and QQ plots. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity had been violated.

4.1.1 Accuracy

Response time For accuracy, we found a significant main effect of error level, $F(3.21, 61.09) = 13.02$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.4$, no significant main effect of target distance, $F(2, 38) = 1.43$, $p = 0.25$, $\eta_p^2 = 0.07$, and a significant interaction between the two factors, $F(4.98, 94.71) = 2.63$, $\boldsymbol{p = 0.02}$, $\eta_p^2 = 0.12$. Figure 4a shows the aggregated response time for the simulated accuracy error levels at different target distances.

Error rate For accuracy, we found a significant main effect of error level, $F(2.99, 56.84) = 4.91$, $\boldsymbol{p = 0.004}$, $\eta_p^2 = 0.2$. We did not find a significant main effect of target distance, $F(2, 38) = 0.41$, $p = 0.66$, $\eta_p^2 = 0.02$, suggesting that the tested target distances did not have a noticeable impact on participants' accuracy of responses. Figure 4b shows the aggregated error rate for the simulated accuracy error levels at different target distances.

4.1.2 Precision

Response time For precision, we did not find a significant effect of error level, $F(2.79, 53.141) = 1.04$, $p = 0.37$, $\eta_p^2 = 0.05$, and target distance, $F(1.61, 30.73) = 2.08$, $p = 0.15$, $\eta_p^2 = 0.09$, on participants' response time, suggesting that the tested target distances and error levels did not noticeably add to the difficulty of the target identification task for this type of error. Figure 4c shows the aggregated response time for the simulated precision error levels at different target distances.

Error rate For precision, we did not find a significant effect of error level, $F(5, 95) = 0.91$, $p = 0.47$, $\eta_p^2 = 0.04$, and target distance, $F(2, 38) = 0.37$, $p = 0.68$, $\eta_p^2 = 0.01$, on participants' error rate, again suggesting that the tested target distances and error levels did not noticeably add to the difficulty of the target identification task for this type of error. Figure 4d shows the aggregated error rate for the simulated precision error levels at different target distances.

4.1.3 Latency

Response time For latency, we found a significant main effect of error level, $F(5, 95) = 16.12$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.45$. We did not find a significant main effect of target distance, $F(1.61, 30.64) = 1.27$, $p = 0.27$, $\eta_p^2 = 0.06$, suggesting that the tested target distances did not noticeably add to the difficulty of the target identification task. Figure 5a shows the aggregated response time for the simulated latency error levels at different target distances.

Error rate For latency, we found a significant main effect of error level, $F(2.96, 56.32) = 14.23$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.42$. We also found a significant main effect of target distance, $F(2, 38) = 5.18$, $\boldsymbol{p = 0.01}$, $\eta_p^2 = 0.21$. Figure 5b shows the aggregated error rate for the simulated latency error levels at different target distances.

4.1.4 Dropout

Response time For dropout, we found a significant main effect of error level, $F(3.41, 64.80) = 26.26$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.58$, no significant main effect of target distance, $F(1.61, 30.74) = 0.61$, $p = 0.51$, $\eta_p^2 = 0.03$, and a significant interaction between the two factors, $F(7.81, 148.56) = 2.68$, $\boldsymbol{p = 0.009}$, $\eta_p^2 = 0.12$. Figure 5c shows the aggregated response time for the simulated dropout error levels at different target distances.

Error rate For dropout, we did not find a significant main effect of error level, $F(2.59, 49.28) = 0.74$, $p = 0.51$, $\eta_p^2 = 0.03$, or target distance, $F(1.39, 26.55) = 1.30$, $p = 0.27$, $\eta_p^2 = 0.06$, on participants' error rate, which suggests that the tested target distances and error levels did not noticeably impact participants' responses. Figure 5d shows the aggregated error rate for the simulated dropout error levels at different target distances.

4.1.5 Error analysis

We analyzed the participants' selections in terms of the selected target compared to the correct target with respect to characteristics of their walking speed and depth. For this, a depth interval of +1 indicates that a participant's selection is one person behind the correct target, and −1 indicates that the selection is one person in front of the correct target. The range of these depth intervals is between ±6 due to the way that the virtual humans were positioned in the crowd as well as where the targets could be positioned (see Sect. 3.2.2). Similarly, as the speed of the virtual humans was randomly assigned from a predefined list, a speed interval of +1 indicates that a participant's selection moved at one speed interval faster than the correct target, and −1 indicates that the selection moved at one speed interval slower than the correct target. These speed intervals had a range between ±8, as there were a total of nine different possible speeds.

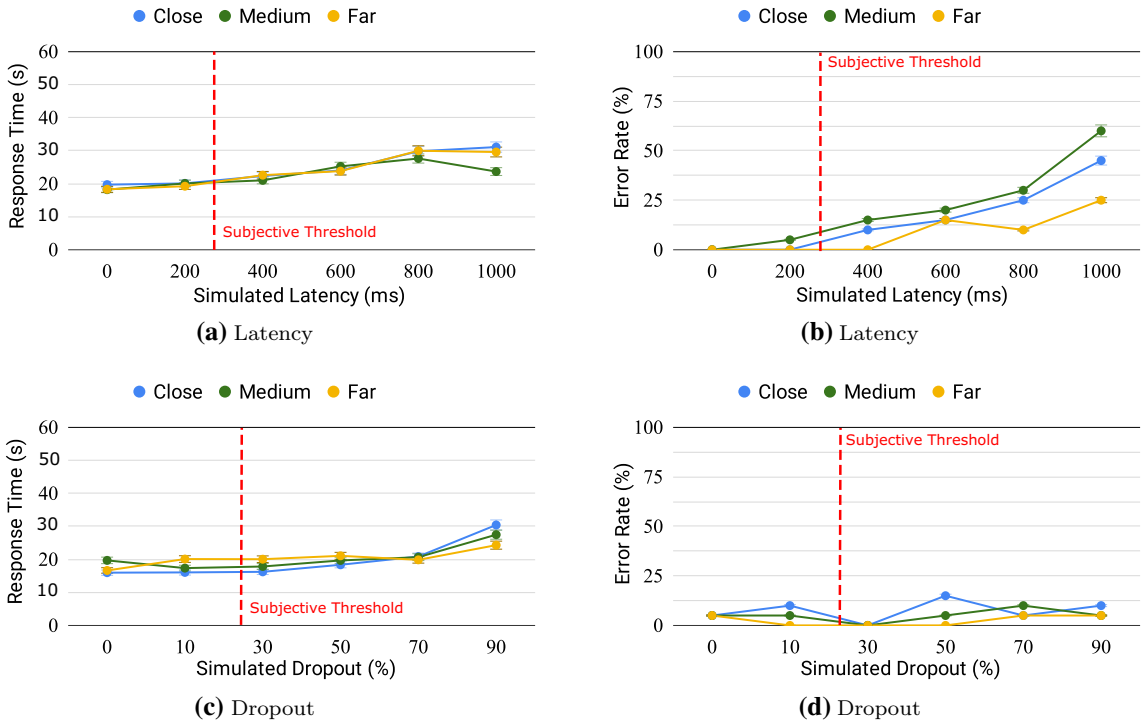


Fig. 5 Performance results related to network errors: The top row (**a**, **b**) shows results for *latency* and the bottom row (**c** , **d**) for *dropout*. The left column (**a**, **c**) shows results for *response time* and the right column (**b**, **d**) for *error rate*

The distributions of these intervals were split into four different categories based on the condition error type, and are shown in Fig. 6.

When analyzing the effects of error type and error level on the average absolute value of speed and depth intervals using the methods described at the beginning of this section, we found several significant main effects. We found that the error type had significant main effects on the depth intervals between the correct target and the participants' selected targets, $F(3,57) = 3.38$, $\boldsymbol{p = 0.024}$, $\eta^2 = 0.15$, indicating that when participants made mistakes in the accuracy and latency conditions, their target choice was significantly farther away from the correct target than it was when mistakes were made in the precision and dropout conditions. We also found that error level had a significant main effect on depth intervals, $F(5, 95) = 7.79$, $\boldsymbol{p < 0.001}$, $\eta^2 = 0.29$, and speed intervals, $F(5, 95) = 4.19$, $\boldsymbol{p = 0.002}$, $\eta^2 = 0.18$, indicating that when participants made mistakes in conditions with higher error levels, then their target choice tended to be significantly farther away and move at a different speed than the correct target.

4.2 Subjective measures

For the subjective questionnaire responses with an ordinal data type, we used non-parametric statistical tests to analyze the responses. We used Wilcoxon signed-ranks tests for the related samples.

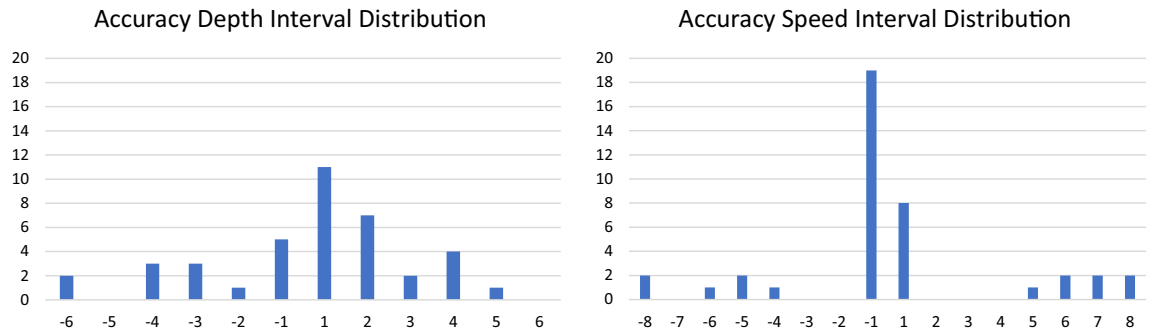
4.2.1 Subjective performance

We analyzed the results for the *Performance* questionnaire (Table 1) with respect to questions SP1, SP2, and SP3. See Table 4.

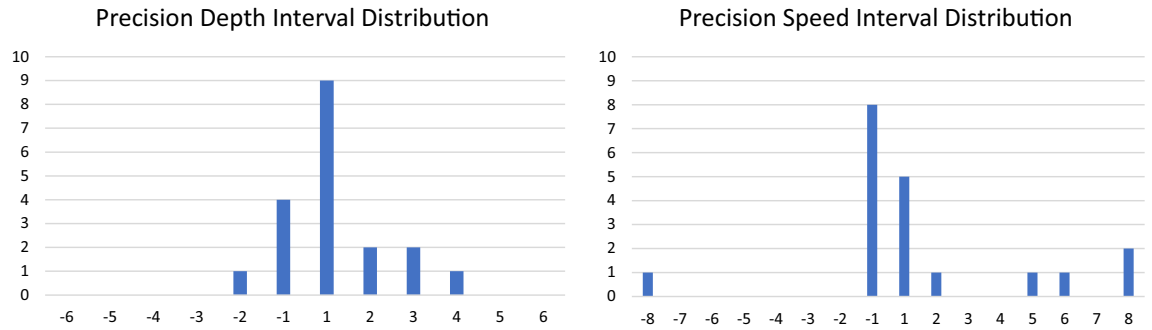
For SP1, we did not find a significant difference among the error types when assessing participants' confidence in their responses, and the similar results for the different error types suggest that they did not have a noticeable impact on self-assessments of their performance.

However, for SP2, we found significant differences between participants' subjective assessment of performance and their actual performance for accuracy, $F(1, 19) = 8.07$, $\boldsymbol{p = 0.01}$, $\eta_p^2 = 0.29$, precision, $F(1, 19) = 20.05$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.51$, and dropout, $F(1, 19) = 29.16$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.6$. We observed a trend for latency, $F(1, 19) = 3.23$, $p = 0.08$, $\eta_p^2 = 0.14$. These results are shown in Fig. 7 and suggest that participants subjectively self-judged their performance as worse than what it actually was.

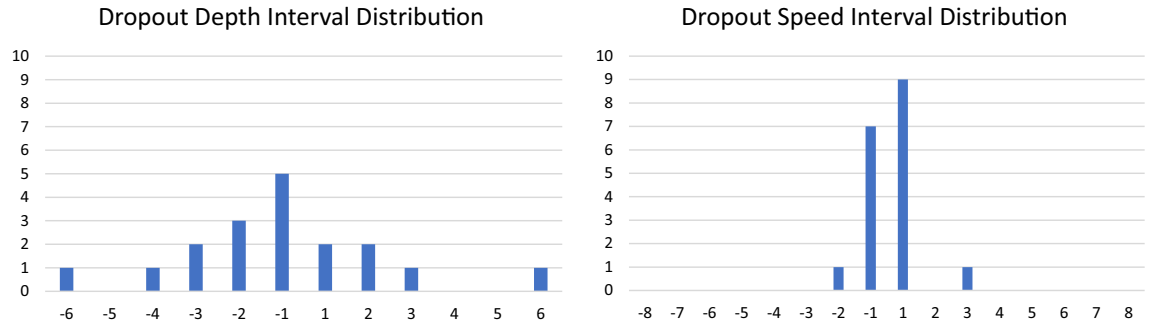
For SP3, we further looked at participants' subjective judgment of what they think constitutes the threshold for an acceptable amount of error. We identified a subjective accuracy threshold, $M = 2.35$ deg, $SD = 1.37$ deg, precision threshold, $M = 1.36$ deg, $SD = 0.64$ deg, latency



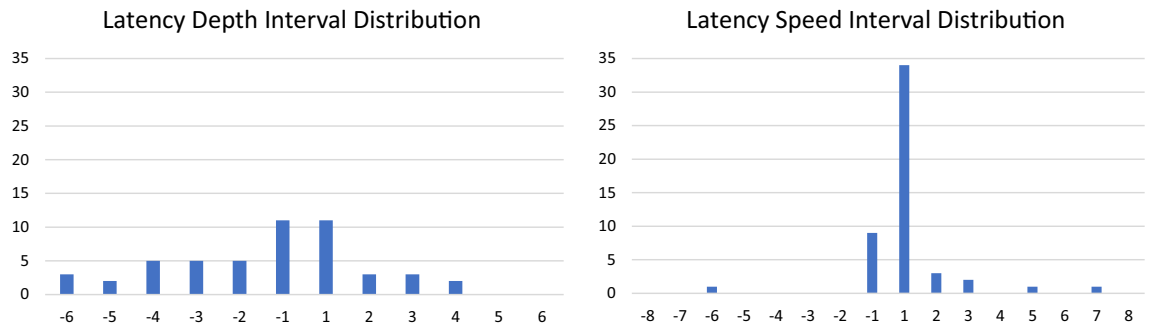
(a) Accuracy Depth and Speed Interval Distributions



(b) Precision Depth and Speed Interval Distributions



(c) Dropout Depth and Speed Interval Distributions



(d) Latency Depth and Speed Interval Distributions

Fig. 6 These figures show the distributions of depth intervals and speed intervals for participants' errors made during the study task. The x-axis show the depth/speed interval as explained in Sect. 4.1.5, and the y-axis shows the frequency of occurrence

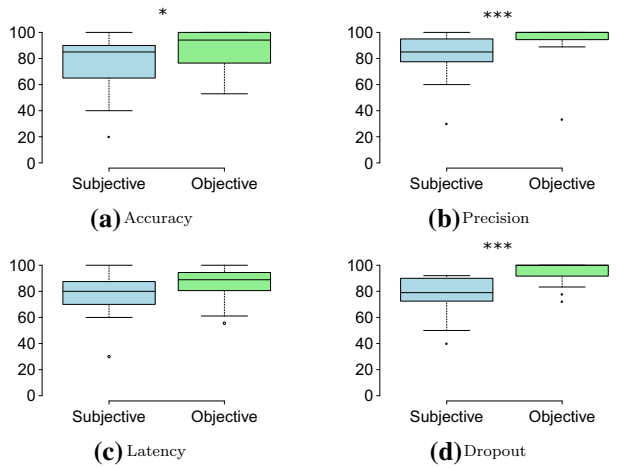


Fig. 7 Comparison of participants' subjective estimates of their performance and their objective performance in terms of correctly identified targets during the experiment for **a** accuracy, **b** precision, **c** latency, and **d** dropout. Statistical significance: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

threshold, $M = 265.9\,\text{ms}$, $SD = 267.7\,\text{ms}$, and dropout threshold, $M = 23.35\%$, $SD = 17.49\%$. We added these subjective thresholds as vertical red lines to the objective measures shown in Figs. 4 and 5. It is interesting to observe that these subjective thresholds seem to be in line with a drop in objective performance for accuracy and latency, while they do not seem to match changes in performance for precision and dropout.

4.2.2 Subjective experience

We analyzed the results for the *Experience* questionnaire (Table 2) with respect to cognitive load questions CL1, CL2, and CL3, as well as ease of use question SUS1, and realism question SE1. See Table 4.

To measure the cognitive load related to each error type, we computed the mean value of CL1, CL2, and CL3. We found significant differences between precision and accuracy, $W = 126$, $Z = 3.01$, $\boldsymbol{p = 0.003}$, between precision and latency, $W = 124$, $Z = 2.9$, $\boldsymbol{p = 0.004}$, and between precision and dropout, $W = 162$, $Z = 2.13$, $\boldsymbol{p = 0.03}$, indicating that the levels of precision error were less mentally demanding. We did not find significant differences for the other comparisons.

For SUS1, we did not find significant differences among the different error types, with overall similar results indicated for ease of use.

For SE1, looking at realism of gaze behavior, we found significant differences between precision and accuracy, $W = 13$, $Z = 2.99$, $\boldsymbol{p = 0.003}$, and a trend between accuracy and latency, $W = 31$, $Z = 1.68$, $p = 0.09$.

4.2.3 Trust in technology

The adapted *Trust in Technology* questionnaire (Table 3) was completed by participants at the end of the experiment. We analyzed the results with respect to questions TT1, TT2, and TT3 focused on participants' perception of the shared gaze interface and questions TT4 and TT5 measuring participants' general view of technology. We calculated the mean values of the questions in each group for TT1–3 ($M = 4.78$, $SD = 1.09$) and TT4–5 ($M = 4.50$, $SD = 1.07$).

We found a significant Pearson correlation between questions focusing on the shared gaze interface and technology, $r = 0.701$, $\boldsymbol{p = 0.001}$. This suggests that participants who have a more trusting outlook towards technology rated the shared gaze interface as better even though 60 out of the 72 trials included some amount of error.

4.3 Qualitative feedback

We conducted short interviews after the experiment to get a better understanding of our participants' impressions of the experience. Most of our participants indicated some level of discomfort associated with wearing the HoloLens for the duration of the experiment. Eleven of our participants reported experiencing slight or moderate amounts of eye strain. Only two of our participants reported experiencing slight amounts of headache or dizziness.

We asked our participants whether or not they observed differences in the gaze behavior of their simulated virtual partner within each block and between any two blocks to gauge how perceptible the simulated differences in error types were. Only three of our participants mentioned that the gaze behaviors within each block were similar and seemed to follow a consistent model. Apart from one participant who mentioned that the gaze behavior seemed similar comparing any two blocks, six participants noted that some blocks had similarities with each other. Interestingly, one of our participants mentioned the impact of the limited field of view of the HoloLens and that they tried to compensate for it by leaning slightly backwards while standing in place.

Some participants remarked on the fact that the gaze rays did not terminate at the body of the target human, which would provide useful depth cues for practical shared gaze environments. As discussed in Sect. 1, we expected this feedback, but such approaches would require accurate real-time information about dynamic scene geometry, which is highly challenging for practical applications, and not considered in the scope of this paper.

Most participants further indicated that they made judgments based on a visual comparison between the movement patterns of the target humans and those of the gaze ray, and that it helped them in completing the task. It would be more challenging to identify a stationary target among a group of

Table 4 Subjective responses for the *Performance* and *Experience* questionnaires

ID	Accuracy	Precision	Latency	Dropout
SP1	5.30 (1.38)	5.55 (1.57)	5.10 (1.37)	5.00 (1.29)
SP2	77 (20.93)	83.40 (16.94)	78.00 (16.17)	77.10 (14.43)
SP3	2.35 (1.37)	1.36 (0.64)	265.9 (267.7)	23.35 (17.49)
CL1–3	3.78 (1.34)	3.00 (1.27)	4.01 (1.25)	3.60 (1.05)
SUS1	4.55 (1.46)	5.05 (1.66)	4.35 (1.53)	4.55 (1.50)
SE1	3.33 (1.46)	4.38 (2.03)	4.00 (1.89)	3.76 (1.75)

We report the means (standard deviations) for the four error conditions

stationary objects. That is, unless one's simulated partner is moving as well, which could provide similar cues as in this study, and would be interesting for future research.

5 Discussion

In this experiment, we observed different effects on participants' performance for error types when error level and target distance were varied. We also noticed the relatively high performance of many of our participants even though the subjective estimates of their performance were lower. Last but not least, we identified subjective thresholds for tolerable error levels. In this section, we discuss our findings and their implications for practitioners.

5.1 Effects of error type on performance and subjective response

To answer our initial research question **RQ1**, based on the differences in the nature of each error, we predicted that **accuracy** and **latency** errors can impact participants' performance more negatively. Indeed, this can be observed in Fig. 7. In the case of our task, the temporal and spatial offset introduced through accuracy and latency errors led to more misinterpretations than **precision** and **dropout** errors. This offset posed a bigger potential for the gaze ray to be on the wrong human for a longer duration of time, causing temporary misinterpretations similar to Holmqvist et al.'s example for accuracy errors [16]. This required participants to rely more on the movement patterns of the human targets and spend more time on target identification.

We also observed that our participants subjectively assessed their performance to be worse than their actual performance, supporting our Hypothesis **H4** and suggesting a response to part of our research question **RQ3**, which is similar to Waltemate et al.'s findings on how participants had a tendency to rather perceive themselves as the cause of error in a visual movement task rather than an introduced system error [40]. We also observed a lower score for cognitive load for precision compared to other error types, partly supporting our Hypothesis **H3b**. Surprisingly, we did not find a difference for SP1 (i.e., confidence in answers) and SUS1 (i.e., ease of use), thus not supporting our Hypotheses **H3a** and **H3c**. We think that the novelty of the interaction, some perceived similarity among blocks, and participants' generally lower self-assessment of performance might have caused them to give similar confidence scores for the different error types, although it is important to note that for all error types the mean confidence was higher than 5 on a scale of 1 to 7.

5.2 Effects of error level on performance and subjective error estimation

We found multiple effects indicating a relative increase in response time and an increase in error rate for simulated error levels for different error types, partly confirming our Hypothesis **H1**. To answer our second research question **RQ2**, we collected subjective estimates of what participants indicated as tolerable thresholds for error levels. Their responses indicate that for our task, thresholds of 2.35 deg for accuracy, 1.36 deg for precision, 265.9 ms for latency, and 23.35% for dropout were acceptable.

We compared these subjective thresholds to the objective performance for **accuracy** and **latency** and found that they indeed seem to be indicating an objective drop in performance, which is supported by our statistical analysis. For less demanding situations, we would go as far as to say that thresholds of 3.5 deg for accuracy and 600 ms for latency are acceptable before the performance drops more drastically. Practitioners may select an appropriate level of performance based on their constraints with respect to the temporal demands and severity of errors of the task at hand, and invest in corresponding eye trackers and network solutions.

In contrast, the subjective thresholds do not seem to match noticeable changes in performance for **precision** and **dropout**. For dropout, we observed most of the significant differences in performance around the highest error value in the tested range (i.e., 90% chance of frame drops). We did not observe a change in performance for precision throughout the tested range, suggesting that our tested values were not large enough to cause any disruptions in our participants' performance. We would like to point out that we chose these

ranges based on reported error levels in the literature and by eye tracker manufacturers (Sect. 3.2), and that it is encouraging for shared gaze applications if these levels are already tolerable.

5.3 Effects of target distance on performance

In contrast to what we had hypothesized in **H2**, there were only a few instances of error types and error levels where we observed a significant decrease in performance for some target distances, and this was not always for the farthest targets. Although there were instances where we saw a significant decrease in performance as the targets were further back such as for accuracy errors, we also observed the opposite effect for some latency and dropout error levels where the response time and error rate was higher for targets at the close and medium distances. We think that this can partly be explained by the limitations in the field of view of the HoloLens HMD used for the experiment. It was more difficult for participants to see the AR gaze ray and the human targets at a close distance since they could easily move out of the augmented field of view.

5.4 Participant selection strategies

By observing the depth and speed distributions for each error type, shown in Fig. 6, we can gain insight into the strategies used by participants when selecting virtual humans from the crowd.

In examining the speed interval distributions, we observe that when users made selection errors, they tended to select virtual humans that were moving at similar speeds to the correct target. This suggests that the primary selection strategy used by participants was judging their speed relative to that of the gaze visualization. This is further supported by the depth distributions, which are in general normally distributed with a higher variance than the speed interval distributions. These normal distributions in the depth interval data would be caused by the randomized nature of assigning depth values to each of the non-targeted virtual humans in the crowd, where high interval values of up to ± 6 are not likely because the depth of the target virtual human is set at one of the three variable depths.

This explanation fits very well for the distributions seen in the dropout and latency conditions. However, the accuracy and precision distributions for the speed intervals show more variance. This increase in variance could be explained by participants making a small portion of their selections based on a factor other than depth or speed, which may be the position of the virtual humans as they approach an endpoint where they switch directions. Several of our participants mentioned using this turning point as a selection strategy due to the patterns in the gaze behavior that it produces. For example, if there is an accuracy or precision error on the gaze visualization, then the correct target and the gaze visualization should switch directions of movement at the same time once the target virtual human hits one of the endpoints of its pacing behavior. Such a strategy would not necessarily work as well for the dropout or latency conditions, however, as for those type of conditions the switch in direction may occur after a certain offset of time. Therefore, it seems likely that this increased variance seen in the accuracy and precision conditions may have come from participants using this location information as part of their selection strategy.

Since speed is an important factor for participants when deciding which virtual human is being targeted, future implementations of shared gaze systems may be able to mitigate the effects of eye tracking errors by assisting users in selecting the correct target by analyzing the speed of entities within the sensors' field of view. For example, if there are several entities moving at similar speeds as the gaze ray, the system could provide an indication as to which of the entities are likely to be confused for each other and an indication as to which target best matches the gaze pattern. Such a feature may help reduce the amount of selection errors that users make, and may be investigated in future work.

5.5 Eliminating error

Because of the impact of latency and accuracy on user performance, future shared gaze systems should take care to mitigate these errors wherever possible. While accuracy based errors may be able to be partially corrected through use of techniques such as bendable rays [37] or motion correlation [39] between points in the users view and user eye movements, latency errors are more difficult to eliminate, especially in applications which involve streaming of large quantities of information. For this reason, future shared gaze systems should prioritize implementations where large amounts of network data, such as dynamic spatial mapping data, can be avoided. For colocated shared experiences, we suggest an implementation similar to what we describe below in section 6 where spatial mapping can be done locally on each user's device. However, for remote shared experiences, we recommend in general prioritizing shared gaze information over other information on the network and employing latency reducing techniques such as those described by Conner and Holden [6] in order to avoid a decrease in user performance.

5.6 Limitations

To fully control for behavioral factors among participants and the trials, in this experiment we chose to use a simulated virtual human partner. Although we did not measure for social presence and co-presence among collaborators, we under-

stand that the social impact of the partner could have been different if a real person was chosen instead or if the virtual human partner initiated a conversation with the participants during the experiment. As a separate experiment, it would be interesting to investigate how the different partners (i.e., real or virtual) and level of interaction with that partner might influence the overall collaborative experience although we expect that the perception of gaze should remain the same for the different partners.

Separately, we chose the same viewing direction for both the participants and the simulated partner, who were standing at a relatively close distance to each other (i.e., one meter). It is important to note that the participants' position relative to the virtual human collaborator, as well as the participants' distances to the targets likely have a significant impact on participants' performance and their overall experience.

It is also possible that the limited field of view of the HoloLens (around 30 degrees horizontal by 17 degrees vertical) may have influenced our results by limiting the number of simultaneously available visual cues which could help the participant in making their decision. It is possible that with increased field of view that we could see improvement in the participants performance, and could even potentially change users' preferences when it comes to gaze visualization. While not investigated here, future work should evaluate the role that such factors play in remote collaboration experiences.

6 Prototype for gaze-based augmented reality collaboration

To facilitate gaze-based interaction between multiple colocated users in AR, we used several software assets and commercial devices to create a prototype platform where each user can be aware of their collaborators' gaze. We implemented gaze rays as used in Sect. 3 as well as other gaze visualizations for users to communicate their focus of attention in a shared space.

Our prototype makes contributions beyond the existing literature by allowing for multiple colocated eye-tracked users wearing AR optical see-through head-mounted displays (AR OST-HMDs) to share eye gaze information. Such a combination is beneficial because it allows for the 3D gaze point to be shared regardless of whether the user is looking at a physical or virtual object, and has the added benefit of not having to share spatial mapping data over the network between users as is the case with prototypes such as CoVAR [33]. As far as we know, ours is the first prototype in which eye gaze information is shared between users of OST-HMDs coupled with mounted eye trackers, and thus extends the work such as that done by Li et al. in which head gaze is coupled with techniques such as the parallel bar technique or double ray technique in order to provide clues as to what object the user is observing along the gaze ray [25].

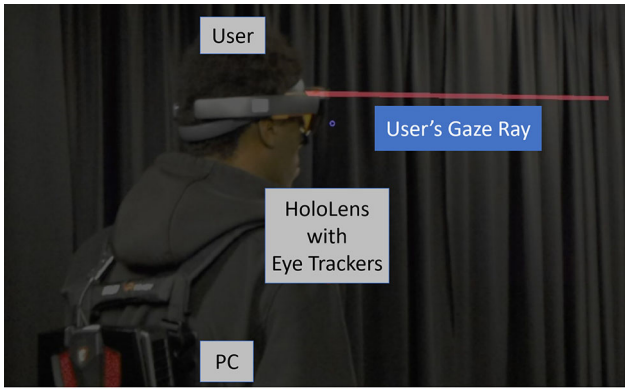


Fig. 8 Annotated screenshot of a user, wearing a backpack computer and a Microsoft HoloLens with the Pupil Labs eye trackers from the point of view of another user, showing the user's gaze ray

We hope that this section will serve as a reference for future work involving shared gaze augmented reality systems by outlining one potential approach in which a prototype system can be implemented.

6.1 Material

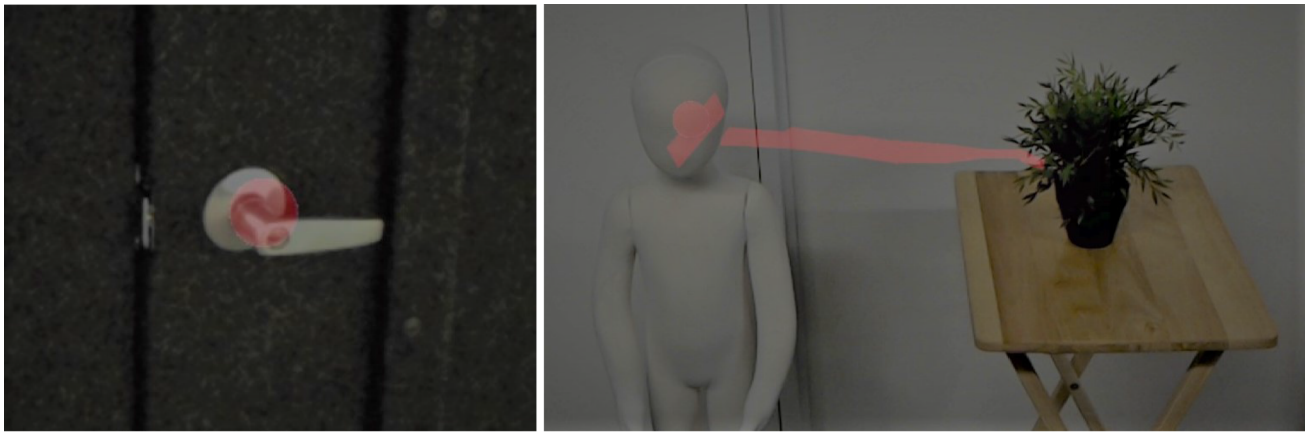
Our prototype platform made use of Microsoft HoloLens HMDs with a Pupil Labs binocular eye tracking add-on.[5] We used Pupil Labs' *hmd-eyes*[6] to calibrate the eye trackers for each user and gain access to 2D and 3D gaze information. The nominal accuracy of the eye trackers reported by the Pupil Labs manufacturer are below 1 deg and 1.5–2.5 deg for the 2D and 3D calibration modes, respectively. We used the 2D approach for calibrating our eye trackers. The eye tracker was tethered to an MSI backpack computer with the following specifications: Intel Core i7-7820HK 2.9 GHz CPU, 16 GB RAM, Nvidia GTX 1070 graphics card, Windows 10 Pro. Figure 8 shows the annotated working prototype.

To create a shared experience between users, we used a combination of the Unity game engine and the Photon Unity Networking (PUN) asset.[7] The PUN package allows for easy network communication between HoloLens devices through Exit Game's Photon Cloud servers. HoloLens clients connect to one of three private "rooms" through the Photon Cloud server, where they can then send and receive information with each other. While we did not measure the end-to-end latency of our prototype system under varying conditions, when using the Photon servers with two simultaneous users, the approximate network latency excluding any system latency

[5] https://pupil-labs.com/products/vr-ar/.

[6] https://github.com/pupil-labs/hmd-eyes.

[7] https://www.photonengine.com/pun.



(a) (b)

Fig. 9 Screenshots of **a** the gaze cursor visualization mode showing a user looking at a door handle and **b** the gaze path visualization mode showing a user looking from the plant to the mannequin

introduced by the eye tracking hardware and 3D rendering in Unity was measured to be 32 ms.

The setup of the shared coordinate system for the prototype involved starting the application from the same position and orientation on each connected HMD instead of utilizing assets such as the holotoolkit's world anchor, which require the application to be deployed to the HoloLens as opposed to running in holographic remoting. While such a method is certainly not ideal for a fully realised implementation, this method allowed for rapid development of prototyping of ideas.

The spatial mapping Unity component was used to allow Unity to access the environment mapping capabilities of the HoloLens, which populated Unity's virtual scene with invisible meshes that represented the user's environment in relation to the user's position. The gaze direction measured by the eye trackers was projected from the user's head position as a Unity ray cast, and the collision point between this ray and the invisible spatial mesh was recorded. This collision point represented the current gaze position of the user, and was sent across the Photon network to be displayed to another user, allowing for real-time sharing of gaze information between users.

6.2 Prototype capabilities

We implemented a gaze ray visualization mode for the prototype, shown in Fig. 8. The color of this gaze ray was easily customized, and is shown in red in the figure. For this type of visualization, the endpoint of the gaze ray was positioned in real time at the user's identified gaze point in the environment, which was computed as the intersection between the gaze ray with the spatial map as described above. From there, the Unity line renderer component was used to draw a line segment between the user's head position and their identified gaze position. This type of gaze visualization gave users access to information about gaze direction and the location of its point of intersection with the environment, which allowed users to identify where other users' attention was focused.

Additionally, we implemented two other gaze visualization modes: gaze cursor and gaze path. For the gaze cursor visualization (shown in Fig. 9a), instead of a line segment connecting a user's head position to their gaze position, solely a circular cursor was drawn at the gaze position without any connecting line segment. This intuitive visualization is similar to using a mouse on a 2D display, however it heavily relies on the depth information gathered from the HoloLens' spatial mapping of the environment, and lacks the directional information that is inherent to the gaze ray visualization mode. This gaze cursor visualization was evaluated along with potential depth based errors that can occur from the depth mapping capabilities of the HoloLens in work by Erickson and Norouzi et al. [9] where it was found that depth errors can also have significant impacts on user performance in shared gaze identification tasks, and that users tend to perform worse with cursor-based visualizations than they do when using gaze ray visualizations due to the lack of the directional cues.

For the gaze path visualization (shown in Fig. 9b), the cursor described above was combined with a Unity line renderer component, so that as the user's gaze position changed, a line was drawn in real time that represented the path of their gaze movement. The gaze position for the most recent n seconds of gaze information is displayed and continuously updated, and the width of the path shrinks over time so that the widest part of the path would be at the user's current gaze position and the smallest part of the path would be where the user was looking 5 s ago.

6.3 Demonstration

We demonstrated our shared gaze prototype at a bi-annual exhibition involving defense researchers and contractors. In our showcased setup, one of the authors wore the eye-tracked HMD and the backpack computer in a room-sized environment. Visitors' could don another HMD and were able to see the experimenter's focus of attention through the different gaze visualizations.

The visitors' feedback was generally very positive, indicating that such shared gaze information can reduce the amount of time spent on conveying positional information to other teammates, reducing the need for hand signals and verbal descriptions of identified target positions, which, depending on the environment, can be difficult to communicate efficiently. With the shared gaze system, a point of interest can be shared in real time, allowing for a reduced reaction time from teammates. Our discussions further indicated a practical demand for using gaze information to indicate targets far beyond the distances tested in Sect. 3, such as at more than a hundred meters in outdoor environments. State-of-the-art eye trackers are currently not sufficiently accurate and precise to pinpoint a target at such distances, limiting their usefulness for such application scenarios.

We further received feedback about the benefits of the gaze path feature in addition to the gaze cursor, as this visualization has the added benefit of being able to instantaneously share information about a moving target among team members. By seeing the shared gaze path, other teammates can quickly learn where the target of interest was recently, in addition to in what direction it is currently moving and its relative speed.

We believe that these applications of shared gaze systems warrant further investigation into the intricacies of such systems and how they can benefit society.

6.4 Prototype limitations

The prototype system relies on gathering environmental information using Unity spatial mapping components, which uses the environmental mapping capabilities of the HoloLens. There are several parameters that can be adjusted when working with this type of component, including the rate at which Unity gathers environment information from the HoloLens, and the number of updates to perform prior to deleting portions of the 3D mesh that have since changed or moved. By varying these values, the correspondence between the generated mesh and the environment is greatly affected as well as the ability of this generated mesh to include dynamic objects.

For the purposes of our prototype, we settled with parameter values that captures static environment information well, but generates artifacts from dynamic objects. Future implementations of shared gaze systems may need to make use of additional depth information to eliminate these difficulties, however the depth information gathered from sensors on the HoloLens or other devices is also subject to various types of errors, as examined by Erickson and Norouzi et al. [9]. Because of this, future work should be carried out to investigate techniques that can mitigate and reduce the effects of these type of errors while the capabilities of the eye tracking and depth sensing hardware continue to improve.

7 Conclusion and future work

In this paper, we investigated the effects of error type, error level, and target distance in an AR shared gaze interface on participants' objective performance and subjective responses through a controlled human-subject study. We designed an experimental scenario inspired by a practical use case of two police officers scanning a crowd of people for a potential threat. In our study, participants were asked to collaborate with a simulated virtual partner and leverage their AR gaze ray to identify a target human among a crowd. We introduced different errors that could impact the data quality presented to the participants either caused by an eye tracker or the network used and measured participants' performance through their response time and error rate in identifying the targets and assessed their subjective experience. We further investigated the current feasibility and methods of implementing a real-time shared gaze system by designing and demonstrating a prototype system.

We identified thresholds for acceptable amounts of error, and our findings suggest that eye tracker accuracy and network latency experienced in current-state shared gaze setups have a noticeable effect on users' performance. In contrast, the tested common ranges of errors for precision and lag were largely acceptable, indicating that these are not a major performance concern for practitioners. We further observed that the field of view of current-state AR HMDs can affect participants' performance with regards to different target distances, and we plan to explore this factor in future work. We also plan to investigate the impact of other visualization techniques, with or without available dynamic scene geometry information, for the gaze cues and how they compare to the gaze ray used in this experiment.

Funding This material includes work supported in part by the Office of Naval Research under Award Number N00014-17-1-2927 (Dr. Peter Squire, Code 34); the National Science Foundation under Collaborative Award Number 1800961 (Dr. Ephraim P. Glinert, IIS); and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting institutions.

References

1. Barz M, Bulling A, Daiber F (2015) Computational modelling and prediction of gaze estimation error for head-mounted eye trackers. DFKI Res Rep 1(1):1–10
2. Bauer M, Kortuem G, Segall Z (1999) "Where are you pointing at?" A study of remote collaboration in a wearable videoconference system. In: Digest of papers. Third international symposium on wearable computers, pp 151–158. IEEE
3. Brennan SE, Chen X, Dickinson CA, Neider MB, Zelinsky GJ (2008) Coordinating cognition: the costs and benefits of shared gaze during collaborative search. Cognition 106(3):1465–1477
4. Brooke J (1996) SUS: a quick and dirty usability scale. Usabil Eval Ind 189(194):4–7
5. Cerrolaza JJ, Villanueva A, Villanueva M, Cabeza R (2012) Error characterization and compensation in eye tracking systems. In: Proceedings of the symposium on eye tracking research and applications, pp 205–208. ACM
6. Conner B, Holden L (1997) Providing a low latency user experience in a high latency application
7. Drewes J, Masson GS, Montagnini A (2012) Shifts in reported gaze position due to changes in pupil size: ground truth and compensation. In: Proceedings of the symposium on eye tracking research and applications, pp 209–212. ACM
8. Ellis SR, Breant F, Manges B, Jacoby R, Adelstein BD (1997) Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: viewing conditions and rendering latency. In: Proceedings of IEEE annual international symposium on virtual reality, pp 138–145. IEEE
9. Erickson A, Norouzi N, Kim K, LaViola JJ Jr, Bruder G, Welch GF (2020) Understanding the effects of depth information in shared gaze augmented reality environments. In: IEEE transactions on visualization and computer graphics
10. Feit AM, Williams S, Toledo A, Paradiso A, Kulkarni H, Kane S, Morris MR (2017) Toward everyday gaze input: accuracy and precision of eye tracking and implications for design. In: Proceedings of the chi conference on human factors in computing systems, pp 1118–1130. ACM
11. Fitzpatrick K, Brewer MA, Turner S (2006) Another look at pedestrian walking speed. Transp Res Rec 1982(1):21–29
12. Geelhoed E, Parker A, Williams DJ, Groen M (2009) Effects of latency on telepresence. Technical report HPL-2009-120, HP Laboratories
13. Gupta K, Lee GA, Billinghurst M (2016) Do you see what I see? The effect of gaze tracking on task space remote collaboration. IEEE Trans Visual Comput Graph 22(11):2413–2422
14. Hall ET (1959) The silent language, vol 948. Anchor Books, New York
15. Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp 139–183. Elsevier, Amsterdam
16. Holmqvist K, Nyström M, Mulvey F (2012) Eye tracker data quality: what it is and how to measure it. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp 45–52. ACM
17. Hornof AJ, Halverson T (2002) Cleaning up systematic error in eye-tracking data by using required fixation locations. Behav Res Methods Instrum Comput 34(4):592–604
18. Jörg S, Normoyle A, Safonova A (2012) How responsiveness affects players' perception in digital games. In: Proceedings of the ACM symposium on applied perception, pp 33–38. ACM
19. Kim K, Billinghurst M, Bruder G, Duh HBL, Welch GF (2018) Revisiting trends in augmented reality research: a review of the 2nd decade of ismar (2008–2017). IEEE Trans Visual Comput Graph (TVCG) 24(11):2947–2962
20. Kim K, Nagendran A, Bailenson J, Welch G (2015) Expectancy violations related to a virtual human's joint gaze behavior in real-virtual human interactions. In: Proceedings of international conference on computer animation and social agents, pp 5–8
21. Kiyokawa K, Takemura H, Yokoya N (1999) A collaboration support technique by integrating a shared virtual reality and a shared augmented reality. In: IEEE proceedings of the international conference on systems, man, and cybernetics (Cat. No. 99CH37028), vol 6, pp 48–53. IEEE
22. Koilias A, Mousas C, Anagnostopoulos CN (2019) The effects of motion artifacts on self-avatar agency. Informatics 6(2):18
23. Langton SR, Watt RJ, Bruce V (2000) Do the eyes have it? Cues to the direction of social attention. Trends Cogn Sci 4(2):50–59
24. Lee C, Bonebrake S, Bowman DA, Höllerer T (2010) The role of latency in the validity of AR simulation. In: IEEE virtual reality conference (VR), pp 11–18
25. Li Y, Lu F, Lages WS, Bowman D (2019) Gaze direction visualization techniques for collaborative wide-area model-free augmented reality. In: Symposium on spatial user interaction, pp 1–11
26. Mcknight DH, Carter M, Thatcher JB, Clay PF (2011) Trust in a specific technology: an investigation of its components and measures. ACM Trans Manag Inf Syst 2(2):12
27. Murray N, Roberts D, Steed A, Sharkey P, Dickerson P, Rae J (2007) An assessment of eye-gaze potential within immersive virtual environments. ACM Trans Multimedia Comput Commun Appl 3(4):17
28. Norouzi N, Erickson A, Kim K, Schubert R, LaViola Jr, JJ, Bruder G, Welch GF (2019) Effects of shared gaze parameters on visual target identification task performance in augmented reality. In: Proceedings of the ACM symposium on spatial user interaction (SUI), pp 12:1–12:11
29. Nyström M, Andersson R, Holmqvist K, Van De Weijer J (2013) The influence of calibration method and eye physiology on eyetracking data quality. Behav Res Methods 45(1):272–288
30. Ooms K, Dupont L, Lapon L, Popelka S (2015) Accuracy and precision of fixation locations recorded with the low-cost eye tribe tracker in different experimental setups. J Eye Move Res 8(1):1–20
31. Pavlovych A, Stuerzlinger W (2011) Target following performance in the presence of latency, jitter, and signal dropouts. In: Proceedings of Graphics Interface. Canadian Human–Computer Communications Society, pp 33–40
32. Piumsomboon T, Day A, Ens B, Lee Y, Lee G, Billinghurst M (2017) Exploring enhancements for remote mixed reality collaboration. In: ACM SIGGRAPH Asia mobile graphics and interactive applications
33. Piumsomboon T, Lee Y, Lee G, Billinghurst M (2017) Covar: a collaborative virtual and augmented reality system for remote collaboration. In: SIGGRAPH Asia 2017 emerging technologies. ACM
34. Piumsomboon T, Lee Y, Lee GA, Dey A, Billinghurst M (2017) Empathic mixed reality: sharing what you feel and interacting with what you see. In: International symposium on ubiquitous virtual reality, pp 38–41. IEEE
35. Ragan E, Wilkes C, Bowman DA, Hollerer T (2009) Simulation of augmented reality systems in purely virtual environments. In: IEEE virtual reality conference, pp 287–288
36. Schoenenberg K (2016) The quality of mediated-conversations under transmission delay. Ph.D. thesis, TU Berlin

37. Steinicke F, Ropinski T, Hinrichs K (2006) Object selection in virtual environments using an improved virtual pointer metaphor. In: Computer vision and graphics. Springer, Berlin, pp 320–326
38. Toothman N, Neff M (2019) The impact of avatar tracking errors on user experience in VR. In: Proceedings of IEEE virtual reality (VR), pp 1–11
39. Velloso E, Carter M, Newn J, Esteves A, Clarke C, Gellersen H (2017) Motion correlation: selecting objects by matching their movement. ACM Trans Comput Hum Interact 24(3):35
40. Waltemate T, Senna I, Hülsmann F, Rohde M, Kopp S, Ernst M, Botsch M (2016) The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In: Proceedings of the ACM conference on virtual reality software and technology, pp 27–35
41. Welch G, Bruder G, Squire P, Schubert R (2019) Anticipating widespread augmented reality: insights from the 2018 AR visioning workshop. Technical report, University of Central Florida and Office of Naval Research
42. Zhang Y, Pfeuffer K, Chong MK, Alexander J, Bulling A, Gellersen H (2017) Look together: using gaze for assisting co-located collaborative search. Pers Ubiquit Comput 21(1):173–186