User Performance and Trust with Network-Connected Augmented Reality Head-Worn Displays in Unstable Network Environments

HIROSHI FURUYA, University of Central Florida, USA
MATTHEW GOTTSACKER, University of Central Florida, USA
ZUBIN CHOUDHARY, University of Central Florida, USA
AUSTIN ERICKSON, KBR Inc., USA
RYAN SCHUBERT, University of Central Florida, USA
GERD BRUDER, University of Central Florida, USA
MICHAEL P. BROWNE, Vision Products, LLC, USA
GREGORY F. WELCH, University of Central Florida, USA

Real-world augmented reality (AR) systems are increasingly network connected in order to succeed in operational spaces dependent on cutting edge sensor and computing technologies, all connected together via the network. This dependency inevitably introduces new sources of error to the AR system through instability in network performance. These errors create familiar problems, such as registration errors, on scales unfamiliar to modern AR systems. The effects of these problems on user performance are difficult to investigate as user performance involves many interrelated factors, some of which are difficult to measure. In this work, we control for user behavior to reduce the complexity of investigating the effects of real-world scale latency and signal dropouts on users' objective task performance and subjective assessments of the AR system. Behavior was controlled by instructing participants to adopt specific behaviors based on different levels of trust and reliance on AR tags and environmental visual cues. We found significant negative effects of both network error sources on subjective assessments of both the AR system and task, as well as objective task completion time and errors committed. Participants' subjective assessments and objective task also varied unexpectedly with increasing network error levels, suggesting that human perception of the effects of network errors are not well calibrated to the actual effects on performance. We also found significant interaction between behavior and network errors on task performance, drawing implications not just for AR system design, but also for AR system operation guidelines.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Laboratory experiments**; **User studies**; *Usability testing*; **Mixed / augmented reality**; HCI theory, concepts and models.

Additional Key Words and Phrases: augmented reality, head-worn display, network, performance, trust, behavior

Authors' Contact Information: Hiroshi Furuya, hiroshi.furuya@ucf.edu, University of Central Florida, Orlando, Florida, USA; Matthew Gottsacker, matthew.gottsacker@ucf.edu, University of Central Florida, Orlando, Florida, USA; Zubin Choudhary, zubin.choudhary@ucf.edu, University of Central Florida, Orlando, Florida, USA; Austin Erickson, austin.erickson@us.kbr.com, KBR Inc., Wright-Patterson AFB, Ohio, USA; Ryan Schubert, ryan.schubert@ucf.edu, University of Central Florida, Orlando, Florida, USA; Gerd Bruder, bruder@ucf.edu, University of Central Florida, Orlando, Florida, USA; Michael P. Browne, m.browne@visionproducts.llc, Vision Products, LLC, Campbell, California, USA; Gregory F. Welch, welch@ucf.edu, University of Central Florida, Orlando, Florida, USA.

ACM 1544-3965/2026/7-ART
https://doi.org/10.1145/3822592

1 Introduction

Augmented reality (AR) technologies have advanced in many aspects in recent years, leading to head-worn displays (HWDs) becoming more attractive for use in many real-world applications, including simulation and training, defense, and urban planning [12, 35]. In these contexts, success in the operational space will be driven in part by the use of distributed, cloud, and edge computing technologies, or the Internet of Things (IoT), to provide critical information to end users who execute the mission [41]. Real-world AR systems will rely on spatial data created and delivered by such networked technologies, greatly increasing the scope of content that can be presented to the user. Networked computing also offers the ability to offload computing requirements to the network, allowing for increased AR system capability while minimizing on-premises resource requirements [3]. At the same time, these dependencies on the network introduce additional risk where degradations in network performance create new sources of error in the AR system. Understanding the effects of such network issues on the performance of an AR system user is necessary to field AR systems in an increasingly network-dependent world. In this paper, we investigate the impact of two factors of these networking-related challenges, network latency and signal interruption ("dropouts" [2]), on human perception and performance with AR systems.

In an AR system that relies on network data to present the real-world location of items of interest, AR cues may lag behind real-world motion as location data arrives late. Further, dropouts may appear as complete stops in cue movement as new data does not arrive to update cue positioning. The exact effects of sub-optimal network performance on AR user performance can vary based on several factors, such as the design of the AR interface, the task at hand, whether users can or are willing to adapt to AR cue positioning and latency errors, and the user's strategy and behavior when using the AR system [18]. These factors can be interrelated and the relationship between AR network errors, user reliance behavior, and ultimate spatial task performance can be complex. For example, trust, defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [15], can affect user reliance behavior which in turn affects task performance. Trust, however, as a latent concept is difficult to directly measure and thus manipulating trust in a system for experiments can also be difficult to accomplish precisely. Therefore, in this work we control for different user reliance behaviors that we would expect to emerge from users with different levels of trust in the AR system. Participants were instructed to adopt pre-defined task completion behaviors modeling different levels of trust and reliance on the information channels available to the user, i.e. AR tag information versus environmental visual information. These behaviors differentiate based on the order in which the information sources are referred to and the priority with which they are used to complete a decision. In doing so, we can evaluate different task completion behaviors, allowing us to better understand the effects of network latency and dropouts on task performance in an AR-assisted visual search-and-selection task.

Accordingly, we conducted two within-subjects experiments (N=20) where we controlled for user reliance behavior to investigate how network errors (i.e., latency and dropouts) affect optical see-through (OST) AR HWD users' objective performance in a 360° search-and-selection task. We also measured users' subjective assessments of the network errors in terms of perceived trust and reliance of the AR system. These experiments support the following findings:

- Effects of AR cue latency and dropouts on participants' objective task performance in the 360° search-and-selection task,
- Interaction effects between implemented behavior and AR latency and dropout errors on task performance, and
- Effects of AR cue latency and dropouts on participants' subjective ratings of reliance and trust in the AR system.

Our results highlight the ability of simple networking errors, occurring in the ranges that we see in difficult real-world networking environments, to induce complex negative effects on objective performance and subjective assessments of the AR system. Our results also demonstrate the potential for different behaviors to mitigate or exacerbate these effects when completing spatial tasks supported by an AR system. Together, this work informs important considerations for the design and operation of AR systems that depend on networking capabilities in the real world.

This paper is structured as follows. Section 2 provides background information in the scope of this paper. Section 3 describes the material and methods we used for our experiments. Section 4 describes the two experiments in which we evaluated the effects of network latency and dropouts. Section 5 concludes our paper.

2 Background

In this section we discuss concepts and prior work that inform our investigation of the relationships between visual search tasks with AR cues, network factors, and reliance and trust relevant to network factors in AR systems, trust in AR contexts, and the search-and-selection task we used.

2.1 User Trust in Augmented Reality Cues During Visual Search Tasks

OST AR HWDs are used in various domains to overlay visual cues onto the real environment. These cues, whether textual or symbolic, annotate physical objects or elements, often providing information that is additional or redundant to what is observable directly in the environment. OST HWDs are particularly advantageous in scenarios where they deliver information about physical objects in a manner that is more efficient and understandable than real-world visual cues [39]. For example, AR annotations that are salient and legible can aid users in locating objects that would otherwise demand extensive effort to find due to small size or being out of view (i.e., occluded) particularly in time-sensitive or unfamiliar settings [6, 26, 34]. Although this information can be technically superfluous, AR imagery offers an alternative channel of information that affords cognitive and performance benefits during task execution. However, evaluating the effectiveness of AR in task contexts is challenging, as both subjective and objective outcomes are influenced by user behavior, which in turn is shaped by their perceptions of the AR system's reliability. Users might exclusively depend on AR cues, combine them with real-world observations (e.g., verify the accuracy of AR annotations against actual environmental elements or vice versa), or disregard the AR cues altogether based on their trust in the technology. In this paper, we propose and assess four distinct behavioral models for executing a 360-degree visual search-and-selection task, delineating variations in the reliance on and trust in the available sources of information.

Factors involved in a user's interaction with the system for a given task include the system's reliability, user attitudes toward such systems, the workload associated with the task, the user's own confidence in their ability to perform the task, and the user's trust in the system [24]. In this context, the reliability of a system — in our study, the reliability of an AR cue — is distinct from a user's reliance, or choice to rely, on that system. Reliability refers to the actual capability of the system, for example what percentage of the time an alarm system performs the correct action. Reliance would then refer to a user's decision to depend on the system for execution of the task.

Trust is yet another related, but different, concept that is drawn from interpersonal trust, or the trust a person has in another person [19]. The exact definition of "trust" in the context of AR systems assisting users in completing spatial tasks is not universally agreed upon. Trust is generally accepted, however, to be a latent concept and one commonly cited definition was offered by Lee and See, "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [15]. A user's trust in and reliance on an AR system are influenced by numerous factors associated with the system, environmental conditions, and the specific task context, often linked to perceived changes in performance efficacy [21, 38, 39]. Reliance and reliability, both constructs related to trust, are typically straightforwardly measured as the percentage of system decisions the human agrees with and the percentage of correct actions over total operation, respectively. Trust, on the other hand, is difficult to directly measure and thus manipulating trust in a system for experiments can also be difficult to accomplish precisely.

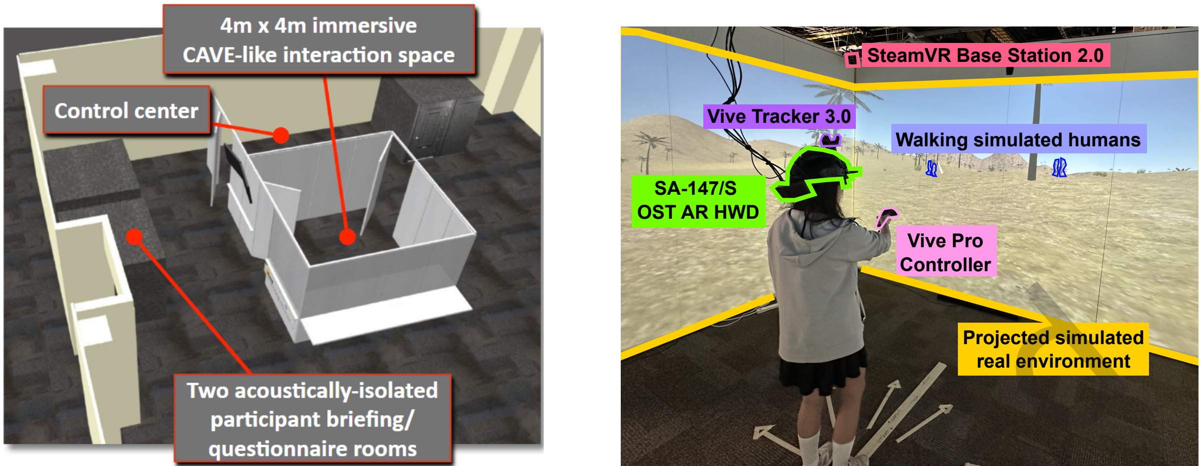
Prior research on reliance and trust informs the expectation that networking factors should have a significant effect on the users' behavior with an AR system that is networked with other users or automated systems to receive task-relevant information (e.g., the GPS position of a moving object to annotate with an AR tag). In these cases, the reliability of various network factors such as latency and dropout errors will impact the user's perceptions of the system. For example, in instances where the system appears unreliable, users may lose trust, leading to reduced usage of the AR system. Such scenarios can result in *undertrusting* the system, where users, doubting the system's reliability, might frequently verify AR-provided information, ultimately impairing performance compared to what could be achieved with a more balanced approach to system trust [24, 29]. On the other hand, excessive confidence in the AR system can lead to *overtrusting* the system, potentially causing more errors than would occur with a more accurately calibrated trust level [22]. Moreover, users' trust and reliance on an AR system can evolve over time, prompting adjustments in their approach to tasks. Additionally, the correlation between users' perceived reliability of the AR system and their actual behavior might not always align, particularly if they are directed to execute tasks in a prescribed manner. While exploring these dynamic behaviors within AR systems is compelling, in this study, we have opted to regulate this variable by specifically directing participants to execute the task employing one of the four predefined behaviors, described in detail in Section 3.2.

2.2 Network Factors in Augmented Reality Systems

AR HWDs are evolving from isolated devices into integrated components of larger, internet-connected systems [31]. This network connectivity offers significant advantages. For example, cloud services can be leveraged to offload computationally intensive tasks, resulting in smoother performance for the user. Additionally, networking can facilitate remote collaboration, enabling AR HWDs to connect with vast IoT networks and access real-time data streams [11, 13]. The quality of future AR experiences will increasingly rely on the seamless transmission of data between the headset and the networked system [40]. Advancements in mobile networking like 5G and the anticipated arrival of 6G by 2030 provide high-speed and high-efficiency infrastructure that supports the continued development of innovative AR applications [8, 10, 16]. These progressive network technology improvements are also reflected in recommendations for network performance for mobile AR applications, including high speeds of at least 50 Mb/s, high reliability of maintaining high speeds at least 95% of the time, low latency of below 10ms, and even high performance while moving at speeds of up to 100km/h [1, 3].

Real-world networking is complex, however, and many more factors than the mobile networking technology by the AR device affect actual network performance. For example, in a worldwide cloud IoT application, a user could expect to experience round trip latency times of more than 1500ms [7], regardless of the networking technology utilized by the end-user's device. In a wireless inertial pose tracking application, real-world packet loss in the WiFi network was measured to be almost 40% in controlled environment without sources of interference [37]. In this paper we investigate the impacts of two aspects of network performance, latency and dropouts, on user perception and performance with AR systems.

Network latency and dropouts can introduce errors that appear as poor AR registration, such as latency and interruptions in tracking. The quality of registration, or the proper positioning of AR cues relative to physical objects and the real world environment, has been long known to directly impact user performance with AR systems [18]. A study utilizing an AR collaborative gaze-sharing application demonstrated that latency and dropout errors induce negative effects on users' objective performance and subjective experience [4]. Another simulated the effects of network and tracking errors on interactions with an avatar in virtual reality, including simulated network errors spiking for 50% of the experiment time [32]. Yet, much of the prior work on the effects



(a) (b)

Fig. 1. In our hybrid reality apparatus, participants were presented with a simulated real space in 360° around them, while they were also wearing an OST HWD that overlaid AR tags over entities within this simulated space: (a) Illustration of the CAVE-like interaction space and control center [27] we used in this work to present a simulated real environment around participants during the experiments. (b) Annotated photo showing a participant completing the experiment, wearing a Vision Products SA-147/S AR HWD in our CAVE-like interaction space, while holding the controller in their dominant hand. The environment showed groups of simulated humans arranged in a circle around the center of the simulation space. Each group of simulated humans walked back and forth while AR tags floating over their heads were presented on the AR HWD.

of AR registration performance on user performance primarily concerns aspects of the AR device itself [9], such as motion-to-photon latency [17] or interruptions in the device's tracking functionality [25]. The increasing reliance on networked systems necessitates a deeper understanding of how network errors, specifically latency and dropouts from network sources as opposed to AR system sources, affect user experience. This paper focuses on investigating the impact of latency and dropout errors, specifically from network sources and not AR system sources, on task performance, trust, and user reliance on the AR system.

3 Material and Methods

In this section, we describe the experimental apparatus, stimuli, measures, and procedure that we used in the two experiments described in Section 4.

3.1 Material

3.1.1 Hybrid Reality Apparatus. For this work we used a 4 m × 4 m square CAVE-like hybrid immersive projection environment [27]. See Figure 1(a) for a diagram of this space. We used four NEC U321H ultra-short throw projectors to project on all four walls. Each projector supported a resolution of 1080p per wall, leading to 7680×1080 total resolution. Participants stood in the center of this interaction space and wore an OST HWD to view AR cues overlaid on top of entities within the projected simulated real environment on the walls. They used a handheld controller to interact with the environment. Figure 1(b) shows a participant standing within this setup. This space allowed us to use an OST HWD to display AR cues overlaid on the real world while granting high levels of control over the alignment between AR cue and target as well as the condition of the environment itself, i.e. weather and lighting.

For the AR stimuli, participants wore a Vision Products SA-147/S OST AR HWD, which uses four OLED microdisplays to support a total resolution of 3840×1200 per eye. The SA-147/S provides a 33° vertical field of view (FOV) and a 143° horizontal FOV, with 53° of binocular overlap. The scene displayed on the projected walls and the AR tags displayed on the HWD were perspective-corrected based on the participant's tracked head pose. We mounted a Vive Tracker 3.0 on top of the HWD to track its pose within the interaction space. Two SteamVR Base Station 2.0 units, mounted in opposite upper corners of the interaction space, tracked the Vive Tracker 3.0 and provided six degrees of freedom head pose information. The Vive Base Station 2.0 units also tracked a Vive Pro handheld controller that provided point-and-click input from the participant. According to the manufacturer, the Vive tracking system has sub-centimeter positional error, sub-degree orientational error, and tracker latency below 30ms [20, 23, 30].

3.1.2 Rendering and Display. We used the Unity engine version 2021.3.2 for rendering. We used the SteamVR Unity plugin to handle the real-time pose and input information from the Vive Tracker 3.0 (for head pose) and the Vive Pro controller. To correctly display imagery on the SA-147/S HWD, we positioned two virtual cameras in the Unity scene based on where the participant's left and right eyes were relative to the Vive Tracker 3.0. This was accomplished by measuring the offsets between the tracker mounting position and approximate locations for a nominal user's left and right eyes. We also manually adjusted the FOV settings on the virtual cameras until they matched the participant's FOV through the AR display. To compensate for the optical distortion characteristics of the SA-147/S HWD, we used a custom shader in Unity, resulting in imagery that appeared undistorted.

3.1.3 Simulated Real Environment. The environment projected on the walls of our immersive projection space was a desert landscape. Ten groups of two simulated humans were distributed evenly around the center of the interaction space at a 10 meter distance, such that each group was located at the same distance from the participant, and the angular distance between each group was the same, as illustrated in Figure 2.

In each group, the two simulated humans patrolled back and forth within the same range of 3 meters distance, at different but constant speeds between 1.5 and 2 meters per second, changing direction every 1.5 to 2 meters. The simulated humans had subtly colored armbands. One of the two simulated humans in each group had a red armband (red team member) and the other a blue armband (blue team member). The order of the colors was assigned randomly. In all conditions, participants saw these colored armbands. Further, in most trials (except for the baseline conditions), participants saw a red diamond AR tag presented via the SA-147/S AR HWD and registered just above the simulated human's head with the red armband. When the simulated human with the red armband patrolled back and forth, the red diamond AR tag matched that human's movement, except for the simulated network latency and dropouts described below. There was no AR tag presented for the simulated human with the blue armband. Hence, depending on the magnitude of the simulated network errors, the simulation ensured that there was uncertainty about which of the two simulated humans the AR tag was registered to.

Game objects in the virtual scene were classified into layers that defined whether they would be presented on the AR display or on the CAVE walls. The culling masks of the cameras for the AR display and CAVE walls were configured such that the terrain and simulated humans in the scene would be presented solely on the walls of the CAVE, while the AR tags above the heads of the simulated humans would be presented solely on the AR display.

An experimenter interface, all four projected wall displays, and the two 3840×1200 pixel video channels for the SA-147/S were driven from a single BOXX APEXX X3 desktop computer via two Nvidia Quadro RTX 6000 graphics cards. HDMI splitters on the four projected wall display outputs allow the wall imagery to be mirrored on four additional displays in the control center area outside the interaction space, where an experimenter could see exactly what imagery was being shown to the participant on each wall. This was useful during the experiments, to be aware of a participant's progress or state while performing the trials, without needing to verbally interact with the participant or to physically go back and forth from the control center area to inside the interaction space.

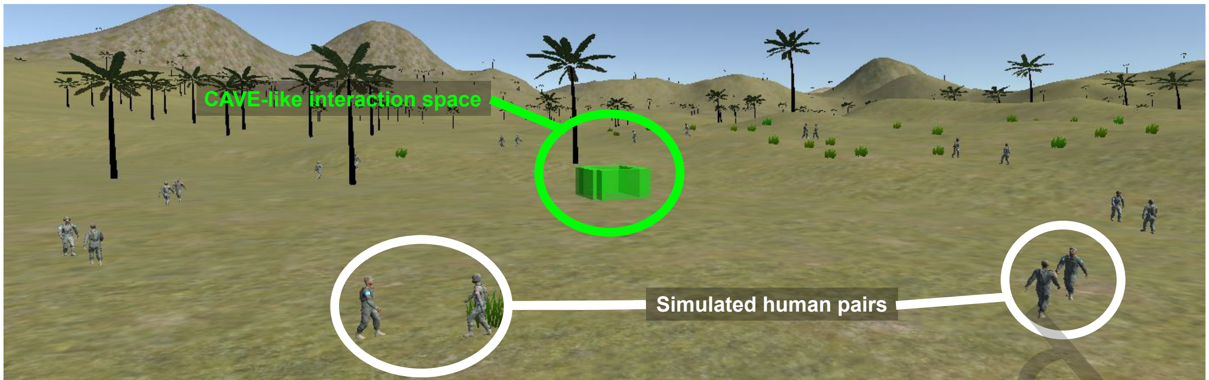


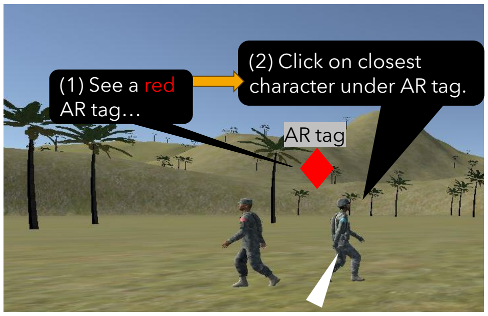
Fig. 2. Screenshot showing the arrangement of the ten groups of two simulated humans that were located in a 10 meter circle relative to the center of the CAVE-like installation (illustrated green). The interior surface of the green walls represents the dimensions and positioning of the CAVE-like space relative to the virtual environment.

3.2 Task and Behaviors

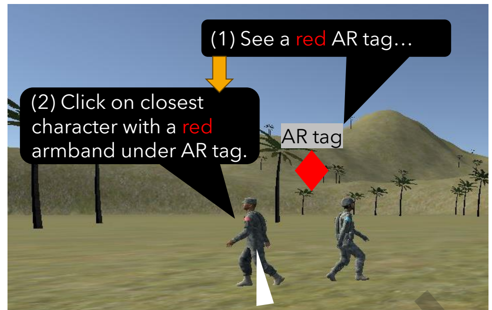
AR-Supported Visual Search-and-Selection Task. The task that participants were asked to perform within the experiment space involved rapidly sweeping around the full 360° surrounding environment, selecting all simulated humans that were marked as red team members, while not selecting any simulated humans marked as blue team members. Throughout the majority of the experiment trials, the simulated humans were identifiable as red or blue team members based on both a colored armband, and the read team members had a red diamond AR tag floating above their head. In baseline trials at the start and end of the experiment, no AR tags were present, requiring participants to rely only on the colored armbands as if they were not wearing an AR display. Participants were timed for each individual trial (one 360° sweep) and were instructed to perform the task as quickly and as accurately as possible. Each trial started with a click on a "start" button and ended with a click on an "end" button. Each trial involved 10 pairs of simulated humans. Each pair of simulated humans featured one blue team member and one red team member. Each trial was repeated twice, once in clockwise direction, and once in counterclockwise direction. Therefore, for each condition the search-and-selection task involved identifying 1 blue team member and 1 red team member out of each pair for 20 pairs of simulated humans. The characteristics of the AR imagery presented to participants in each trial were varied according to the experimental conditions. Additionally, participants were instructed to adopt one of four behaviors that represented different levels of reliance on AR.

Behaviors. We considered four behaviors for completing the AR-supported visual search and selection task in the scope of this study. We identified these four behaviors based on different levels of reliance on and trust in the two information channels (AR tags presented on the HWD vs. colored armbands in the simulated real environment) that are provided to participants, i.e., which information channel participants mainly rely on, and whether they verify that information with the other information channel:

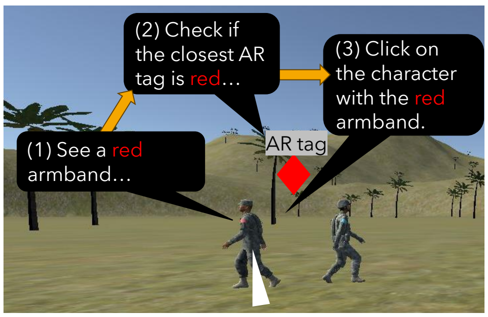
- **AR-Only**: Participants completed the task by only relying on the AR tags while disregarding the simulated real scene armbands.
- **AR-First**: Participants completed the task by first looking at the AR tags, then confirming their impression of the simulated humans by double checking the appearance of their armband.
- **Real-First**: Participants completed the task by first looking at the armbands on the simulated humans' arms, then confirming their impression by double-checking the AR tag.



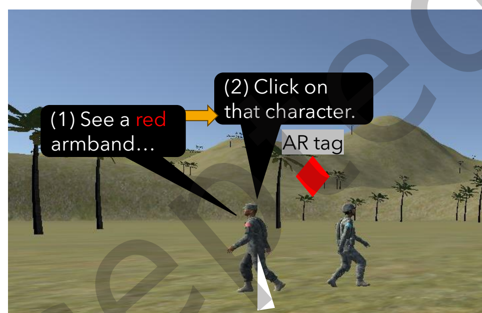
(a) AR-Only



(b) AR-First



(c) Real-First



(d) Real-Only

Fig. 3. Illustration of the four considered behaviors that correspond to different levels of trust and reliance on AR. We instructed participants to use these behaviors when making their decisions about which simulated human to select when completing the AR-supported visual search-and-selection task. (a) AR-Only: Participants only looked at the AR tag when deciding whether to select them. (b) AR-First: Participants first looked at the AR tag and then looked at the simulated human nearby to confirm their decision before selecting. (c) Real-First: Participants first looked at the simulated human's armbands and then looked at the AR tag to confirm their decision before selecting. (d) Real-Only: Participants only looked at the simulated human's armband when deciding whether to select them.

- **Real-Only**: Participants completed the visual search-and-selection task by only relying on the simulated real scene while disregarding the AR tags, i.e., participants identified targets based only on the color of the armbands on the simulated humans.

Depending on a condition's assigned behavior (See Sections 4.1.3 and 4.2.3), we expected to see differences in participant task performance (time and errors, see below). In particular, we expected to see increased task performance for behaviors with higher reliance on and trust in AR (if warranted by the simulated AR network performance). Further, depending on the simulated AR network errors, we expected that participants would subjectively indicate that they would lose trust and lower their reliance on AR in situations when the AR system performance did not seem to warrant trust.

During experimental trials, participants were shown a graphic corresponding to the desired behavior for the trial, seen in Figure 3. A recorded voice instruction was also played while the graphic was displayed to provide two modes of instruction. These instructions were provided in addition to the familiarization period participants were given (see Section 3.3.3).

3.3 Measures and Procedure

We collected both objective and subjective data from participants in our experiments.

3.3.1 Objective Data. We recorded the duration it took participants to conduct a complete 360° sweep of the environment in each trial. This involved measuring the time from when they initiated the trial by clicking the "start" button until they indicated completion by clicking the "end" button. Additionally, we tracked the number of false-negative instances (type II errors), where participants failed to identify simulated humans marked in red, and the number of false-positive instances (type I errors), where participants incorrectly selected simulated humans marked in blue.

3.3.2 Subjective Data. We gathered subjective data from participants using questionnaires completed on a laptop immediately after they finished the search-and-selection trials. This occurred while they were still immersed in the hybrid simulation environment, with simulated humans and AR tags visible in the FOV. The questionnaires utilized in this study were as follows:

Trust of Automated Systems Test (TOAST). This questionnaire [36] provides an effective assessment of whether users of a system believe they understand the system and feel confident in its performance. Participants read each of the ten statements and indicated on a 7-point Likert scale the extent to which they disagree (1) or agree (7) with each statement. This questionnaire is built around the following two subscales, and is scored separately. To calculate the subscale scores, the average of the responses to the items in that subscale is computed, based on the method proposed by Wojton et al. [36]:

- *System Understanding:* Higher understanding scores indicate that the participant has a high confidence that their trust in the system is well calibrated.
- *System Performance:* Higher performance scores indicate that the participant trusts the system to help them perform their task.

Single Item Questionnaires. We asked participants to further rate their perception of the different experimental conditions with respect to the following 7-point scales using a single item each:

- *Trust*: On a scale from 1 (not trustworthy) to 7 (very trustworthy), how much would you trust the AR system?
- *Reliance*: On a scale from 1 (cannot rely on) to 7 (very reliable), how much would you rely on the AR system?
- *Confidence*: How would you rate your confidence in your performance (1=low, 7=high)?
- *Difficulty*: How would you rate the difficulty of the task (1=low, 7=high)?
- *Advantageous*: On a scale from 1 (disadvantageous) to 7 (advantageous), how would you rate the AR system?

These questionnaires were administered with regards to the network factors only. Our work primarily aimed to investigate the effects of network latency and dropouts on task performance; the behaviors were implemented to control for subjective factors that could affect performance. Therefore, when answering the above questionnaires we instructed participants to consider the different network factor levels and did not analyze results by behavior (see Sections 4.1.4 and 4.2.4.

3.3.3 Procedure. Upon arrival, participants were provided with a hard copy of the informed consent document, which they read and signed along with the experimenter. Participants were guided to the CAVE-like simulation environment where verbal consent was obtained to place the SA-147/S AR HWD on their heads.

The SA-147/S AR HWD was fitted on the participants' heads using the following procedure: (1) Fit onto the head using knobs to tighten or loosen the back, sides, and top. (2) Rotate eyepieces down into position in front of the participant's face. (3) Adjust eyepiece positioning by adjusting the angle (pitch) of the eyepieces to be vertical, setting the height of the eyepieces to align with the participants' eyes, adjusting the eyepieces backwards toward the participant's face to ensure closeness without contact, and independently adjusting each eyepiece laterally while the participant closed the other eye to account for their interpupillary distance.

During the HWD adjustment process, a rectangular 2D calibration image was displayed to ensure that the final adjustments permitted participants to see the AR imagery sharply and see all four edges of the FOV. Once the HWD was calibrated, we performed a boresight calibration by instructing participants to gaze straight ahead at a simulated horizon line and applying a one-time tracker pitch rotation offset.

Following the completion of HWD adjustment and tracker calibration, the experimenter briefed participants on the search-and-selection task and experimental conditions. Participants were then given the opportunity to practice the task until they felt comfortable to continue.

The experiment itself comprised baseline tasks conducted before and after a series of tasks incorporating the experimental variables. Following the completion of the selection tasks, the HWD was removed and participants were provided with questionnaires corresponding to each experienced latency and dropout level. In these questionnaires participants were provided with graphics similar to Figures 4 and 8 to help remind them of the different conditions. In addition, each trial was accompanied with a label identifying the current latency or dropout level to ensure participants could differentiate between the different network factor levels. Participants then completed a demographics questionnaire, received a debriefing, and were compensated monetarily.

4 Experiments on AR Network Factors

In this section, we describe the two experiments in which we evaluated the network factors *latency* and *dropouts* using the material and methods described in Section 3. These experiments were approved by the institutional review board (IRB) of our university and the United States Air Force Human Research Protection Program (USAF HRPP).

4.1 Experiment E1: Latency

Here we describe the first of the two experiments, evaluating the effects of network latency. See Section 4.2 for the description of the second experiment evaluating the effects of network dropouts.

4.1.1 Participants. We performed an apriori power analysis in G*Power [5] to estimate an appropriate sample size. We calculated the sample size for a repeated measures, within factors ANOVA for 12 groups and 20 measurements, large effect size of $\eta_p^2 = 0.2$ and standard values of $\alpha = 0.05$ and $1 - \beta = 0.8$. This yielded a minimum sample size of 16 participants.

Based on this approximation we recruited 20 participants from our university community for this experiment, and we obtained participants' demographics using ACM's DIE Demographics Questionnaire. Our participants included 15 males, and 5 females, with ages between 21 and 42, $M = 25.2$, $SD = 5.6$. None of the participants reported any motor or cognitive disabilities. All of the participants had normal or corrected-to-normal vision and none of them reported any visual or vestibular disorders, such as color or night blindness, dyschromatopsia, or a displacement of balance. The participants were either students or non-student members of our university community who responded to open calls for participation and received monetary compensation for their participation. The experiment took participants on average 30 minutes to complete.

4.1.2 Material. In this experiment, we used the task stimuli described in Section 3.1.3. We simulated different levels of network latency in the AR stimuli. Specifically, we applied latency to the movement of the red diamond AR tags that were registered with the humans with the red armbands. The effect of this latency was that the AR tags lagged behind the movement of these humans with the red armbands. As there were two humans moving within a limited distance from each other and on similar motion paths, latency led to uncertainties among participants as to which of the two humans the AR tag was registered with.

We decided to consider latency levels of magnitudes that are common in networked systems [7], i.e., beyond that are common for regular AR tracking and rendering processes. Specifically, we decided to test latency levels of +0 ms, +750 ms and +1500 ms. Compared to these simulated network latency levels, the base tracking and rendering latency of our AR system was negligible. See Figure 4 for an illustration.

We formally controlled AR network latency in this experiment by using AR tag positions that were pre-calculated. In this system, on each frame and for each AR tag we (a) computed the AR tag position as if there were no latency and tagged it with the current system timestamp, and (b) inserted the timestamp-position pair into a list ordered by timestamp. We used this list to simulate AR system latency by performing the following on each frame for each AR tag: (1) We searched for the timestamp-position pair whose difference between key timestamp and current system timestamp was closest to the desired latency. (2) We set the AR tag position to the position in the resulting timestamp-position pair.

It is important to note that we simulated and introduced latency as *network* latency but not as *tracking* or *rendering* latencies. By that we mean that the positions of the AR tags in the environment were delayed due to network latency, but the general head tracking and rendering performance with respect to the AR HWD were not slowed down or delayed. The rendering on the AR HWD was updated in real time based on the high performance of our AR head-tracking system when participants rotated their heads in the experimental space. To ensure that system latency would not affect our results, we measured the motion-to-photon latency of our AR system to be 50ms, which we found to be very difficult to perceive compared to the simulated network latency levels of 750ms and 1500ms. During the experiment, no participants commented or noted significant baseline latency in the system.

4.1.3 Study Design. To investigate trust in AR with respect to latency and different behaviors to perform the 360° search-and-selection task (see Section 3.2), we used a within-subjects design for our study with the following network factor, baseline, and behaviors.

- **Latency (3 levels):** We simulated the three different network latency levels of high (+1500 ms), medium (+750 ms), and low (+0 ms).
- **Baseline (pre/post)**: We included two baseline tests (before and after the experimental trials), in which participants wore the AR HWD, but it was turned off, i.e., participants had to base their decisions entirely on the simulated real environment. We included these baselines mainly as a sanity check with respect to learning effects during the study.
- **Behaviors (4 levels)**: Additionally to the tracking factor listed above, we modeled the four considered behaviors for completing the AR-supported visual search-and-selection task in the scope of this study, which correspond to different levels of reliance on the AR system for the task together with lesser or higher trust in AR.

Behaviors were blocked and randomized, with the order of latency conditions randomized within each behavior block. Each condition was tested twice, once while rotating clockwise and once in counterclockwise direction to avoid that our participants get entangled in the cables that were suspended from the ceiling.

4.1.4 Results. We analyzed the responses with repeated-measures analyses of variance (RM-ANOVAs) and Tukey multiple comparisons with Bonferroni correction at the 5% significance level. We confirmed 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 was not supported.

We found no significant difference between the clockwise and counterclockwise trials as well as the pre and post baselines, so we pooled the responses with respect to our analysis.

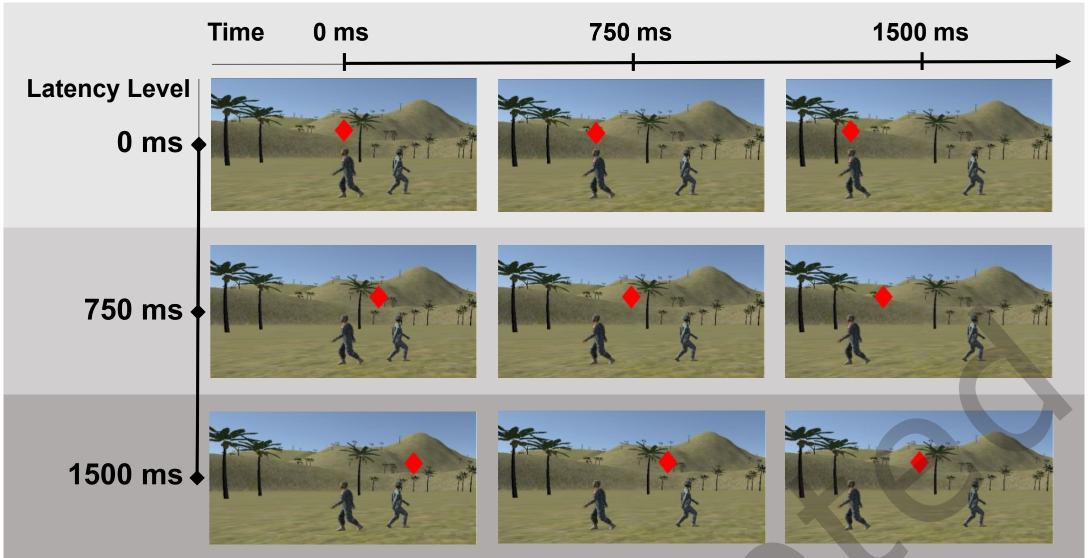


Fig. 4. Illustration of network latency in Experiment E1. Each row illustrates the positions of the red-team simulated human (left human) and their corresponding AR tag at different times. For illustration purposes, the position of the blue-team simulated human (right human) is fixed in these images, though they were also moving in the actual experiment. The different network latencies are characterized by the AR tags lagging behind the position of the red-team simulated human when they are moving.

Objective Data. The descriptive statistics for the elapsed trial times are shown in Figures 5 and 6, and the statistical test results are shown in Table 1.

Our results show significant main effects for the latency levels and behaviors on elapsed time, but no interaction effect between the two factors. Specifically, we found that elapsed times were significantly highest (= worst) for the 1500 ms latency, followed by the 750 ms latency, and the 0 ms latency. Further, we found that Real-First resulted in the significantly highest (= worst) elapsed time compared to the Real-Only and AR-First behaviors.

Our results further show significant main effects as well as interaction effects for the latency levels and behaviors on type I and type II errors. Specifically, we found that both type I and type II errors were significantly lowest (= best) for the 0 ms latency condition compared to the 750 ms and 1500 ms latency conditions. Moreover, both type I and type II errors were significantly highest (= worst) in the AR-Only condition compared to all other conditions. In addition, since testing each condition involved 40 simulated humans (20 pairs, as discussed in Section 3.2), we find that mean error rates approaching 10% for each of Type I and Type II for the high latency and AR-Only behavior condition. In contrast, for all other behavior conditions we see error rates below roughly $1/40 = 2.5\%$ for each Type I and Type II errors.

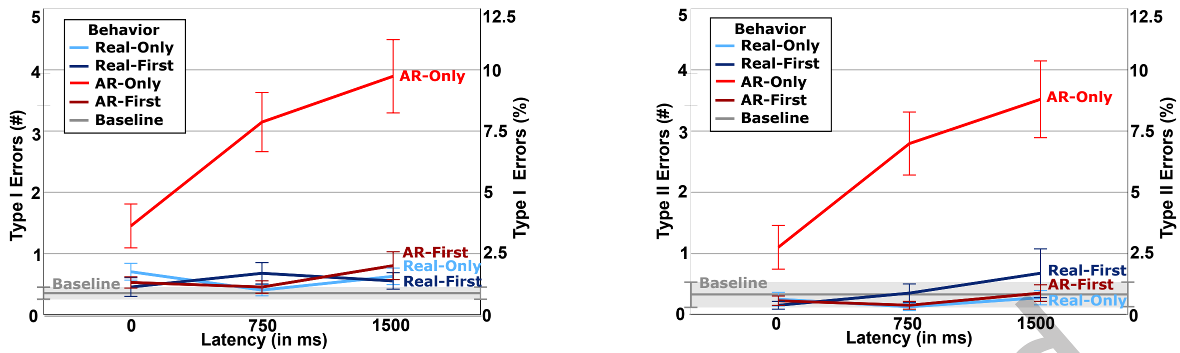
Learning Effect Analysis. Learning effects were examined using linear mixed-effects models with trial index treated as a continuous within-participant predictor and participant modeled as a random intercept. For all analyses, the same model structure was used: `DV ~ Trial + (1 | PID)`. Models were fit using maximum likelihood estimation. A linear mixed-effects model predicting task completion time revealed a significant effect of trial index ($\beta = -0.112$, SE $= 0.019$, $t = -5.74$, $p < .001$), indicating a reliable decrease in completion time across successive trials. The random intercept variance for participant was $\sigma^2 = 21.01$. For Type I error rates, trial index was not a significant predictor ($\beta = -0.002$, SE $= 0.007$, $t = -0.33$, $p = .75$), indicating no systematic change

Table 1. Experiment E1: Statistical test results for the latency network factor and objective performance.

Measures	**RM-ANOVA**	**Factors**	$\mathbf{df}_G$	$\mathbf{df}_E$	$\boldsymbol{F}$	$\boldsymbol{p}$	η_p^2	**Pairwise Comparisons**
Elapsed Time	Two-way	Latency	1.42	26.88	7.89	**0.005**	0.29	All pairs $\boldsymbol{p}$<**0.05**
		Behavior	3	57	4.07	**0.035**	0.14	All $\boldsymbol{p}$<**0.05**: (Real-First > Real-Only), (Real-First > AR-First)
		Latency * Behavior	3.08	58.60	2.27	0.088	0.11	N/A
Type I Errors	Two-way	Latency	2	38	10.18	< **0.001**	0.35	All $\boldsymbol{p}$<**0.05**: (750ms Latency > 0ms Latency), (1500ms Latency > 0ms Latency)
		Behavior	1.24	23.50	33.51	< **0.001**	0.64	All $\boldsymbol{p}$<**0.05**: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First)
		Latency * Behavior	2.42	46.06	9.03	< **0.001**	0.32	All $\boldsymbol{p}$<**0.05**: For 0ms Latency: None For 750ms Latency: (Real-First > AR-First) For 1500ms Latency: (Real-First > Real-Only) For Real-Only behavior: None For Real-First behavior: None For AR-Only behavior: (750ms Latency > 0ms Latency), (1500ms Latency > 0ms Latency) For AR-First behavior: None
Type II Errors	Two-way	Latency	1.44	27.38	11.99	< **0.001**	0.39	All $\boldsymbol{p}$<**0.05**: (750ms Latency > 0ms Latency), (1500ms Latency > 0ms Latency)
		Behavior	1.33	25.23	30.87	< **0.001**	0.62	All $\boldsymbol{p}$<**0.05**: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First)
		Latency * Behavior	2.58	48.96	6.19	**0.002**	0.25	All $\boldsymbol{p}$<**0.05**: For 0ms Latency: None For 750ms Latency: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First) For 1500ms Latency: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First) For Real-Only behavior: None For Real-First behavior: None For AR-Only behavior: (750ms Latency > 0ms Latency), (1500ms Latency > 0ms Latency) For AR-First behavior: None

in false positive errors over trials. The participant-level random intercept variance was $\sigma^2 = 0.22$. Trial index did not significantly predict Type II error rates ($\beta = -0.008$, SE = 0.010, $t = -0.86$, $p = .39$). Participant-level variance was $\sigma^2 = 0.26$, indicating stable miss rates across repeated trials.

Comparison to Baseline. To examine how each behavioral condition differed from baseline, we conducted within-participant comparisons using paired two-sided tests. For elapsed time, performance differed significantly from baseline only in the Real-First condition, which resulted in longer completion times relative to baseline ($t(19) = 3.41$, $p = .003$, $d = 0.72$). No significant differences in elapsed time relative to baseline were observed for AR-First ($p = .278$), AR-Only ($p = .096$), or Real-Only ($p = .818$). For Type I errors, error rates were significantly higher than baseline for AR-First ($t(19) = 2.23$, $p = .038$, $d = 0.51$) and AR-Only ($t(19) = 6.81$, $p < .001$, $d = 2.12$). Differences between baseline and Real-First ($p = .193$) or Real-Only ($p = .094$) were not statistically reliable. For Type II errors, a significant increase relative to baseline was observed only for AR-Only ($t(19) = 5.61$, $p < .001$, $d = 1.57$). No significant baseline differences were found for AR-First ($p = .668$), Real-First ($p = .757$), or Real-Only ($p = .615$).



(a) Type I Errors

(b) Type II Errors

Fig. 5. Experiment E1: Results for (a) Type I errors and (b) Type II errors for the latency trials for the four behaviors and baseline condition. Lower is better. The error bars show the standard error.

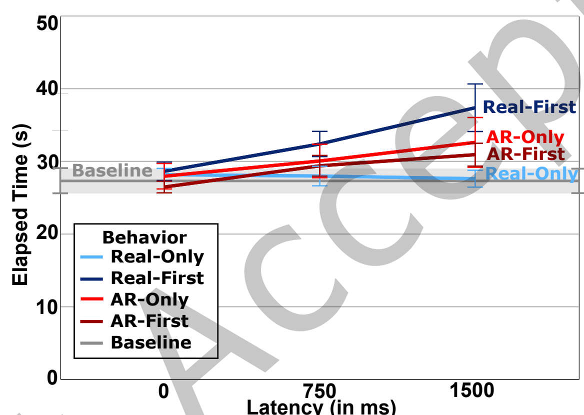


Fig. 6. Experiment E1: Results for elapsed time for the latency trials for the four behaviors and baseline condition. Lower is better. The error bars show the standard error.

Subjective Data. The descriptive statistics for the subjective responses are shown in Figure 7, and the statistical test results are shown in Table 2.

Our results show significant main effects of the latency levels on both subscales of the TOAST questionnaire as well as our single-item questionnaires for Trust, Reliance, Confidence, and Advantageous, but not for Difficulty. All pairwise comparisons were significant (except for Difficulty). Overall, participants exhibited significantly less trust, reliance, and confidence in the AR system and they felt that it was less advantageous the higher the network latency was.

4.1.5 Discussion.

Network Latency. Our results show that network latency had a significant impact on participants' task performance in terms of both completion time and errors for our tested search-and-selection task. It took participants less time to make their selection decisions and they made fewer false-positive and false-negative selection

Table 2. Experiment E1: Statistical test results for the latency network factor and subjective assessment.

Measures	RM-ANOVA	Factors	$\mathbf{df}_G$	$\mathbf{df}_E$	F	p	η_p^2	Pairwise Comparisons
TOAST System Understanding	One-way	Latency	1.47	27.92	28.64	**<0.001**	0.60	All pairs ***p*<0.05**
TOAST System Performance	One-way	Latency	1.54	29.35	86.18	**<0.001**	0.82	All pairs ***p*<0.05**
Trust	One-way	Latency	1.52	28.97	76.69	**<0.001**	0.80	All pairs ***p*<0.05**
Reliance	One-way	Latency	2	38	88.31	**<0.001**	0.82	All pairs ***p*<0.05**
Confidence	One-way	Latency	1.13	21.51	37.14	**<0.001**	0.66	All pairs ***p*<0.05**
Difficulty	One-way	Latency	1.09	20.88	1.46	0.24	0.07	N/A
Advantageous	One-way	Latency	2	38	82.77	**<0.001**	0.81	All pairs ***p*<0.05**

errors when there was no simulated network latency. This result was not unexpected because latency introduces ambiguity when determining which moving human was the selection target, and participants needed to take more time to determine the correct human to select. However, it is interesting to see that the time required to complete the task seemed to increase linearly with the network latency (see Figure 6).

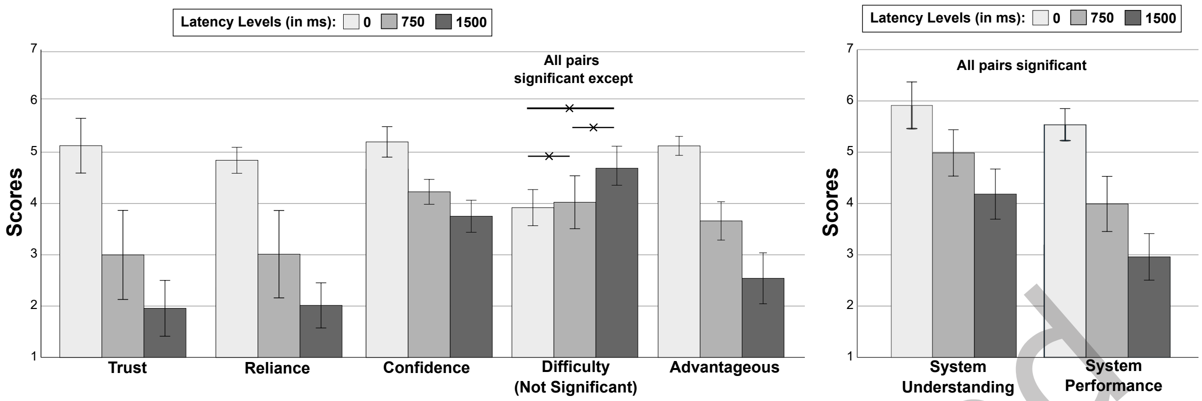
It should be noted that the two simulated humans that participants had to decide between were moving in similar patterns (walking back and forth between two points), so the red diamond AR tag followed similar paths for both humans (red and blue armband). If the characters had moved in different and less predictable patterns, it might have been easier for participants to disambiguate AR cues that were affected by latency. In other words, our results are conservative in the sense that network latency would likely have less pronounced effects in less demanding task scenarios.

Our experiment further showed that network latency significantly impacted participants' trust in the AR system. As latency increased, participants' trust in the AR system decreased. As participants had to spend more time and effort to interpret the information from the AR system with additional latency, and the AR information became difficult to interpret, it affected participants' trust in the AR system. Apart from trust, our results further show similar significant effects for participants' reliance on AR, confidence in their performance, and how advantageous they rated the AR system.

Behaviors. The behaviors participants employed for the visual search-and-selection task had a significant main effect on participants' task completion time and errors. Specifically, Real-First was slower than AR-First and Real-Only, and AR-Only resulted in more errors than any other behavior.

Looking at the AR-Only condition, in which participants relied only on the AR tags to make their selections, the number of type I and II errors made by participants is very noticeable. However, it is further noticeable that these errors all but disappeared when participants used a behavior by which cues from both information channels (AR tags vs. armbands) were integrated and cross-verified (i.e., AR-First or Real-First). One may expect this integration to result in a significant increase in elapsed time, but we found no significant difference between AR-Only and the other conditions. However, we found a significant difference between AR-First and Real-First in elapsed time, showing that between these two behaviors that integrate both information channels, AR-First seems preferable over Real-First.

If faced with the decision between one of these behaviors, a decision which will likely be affected by participants' level of trust in and reliance on the system, we see the pattern that the AR-First behavior is generally preferable over the Real-First behavior, and both are preferable over the AR-Only behavior. Both the AR-First and Real-First behaviors reduce errors, but the Real-First behavior was the slower (= worse) condition compared to the AR-First behaviors. It is further noticeable that with increasing AR network latency levels, there may come the point



(a) Single-Item Questionnaires (b) TOAST Questionnaire

Fig. 7. Experiment E1: Subjective data results for the three levels of latency with (a) Single-Item Questionnaires and (b) TOAST Questionnaire. Higher is better (except for Difficulty). The horizontal lines with crosses in the middle indicate those pairs that were *not* significant ($p > 0.05$). The error bars show the standard error.

for applications where it may increase performance to disregard the AR information entirely and rely only on real-world cues.

Interactions. We only found significant interactions between latency and behavior for type I and II errors. These interactions primarily center around the AR-Only behavior. These demonstrate that when participants only follow the AR cue without confirming the correctness of the current cue placement, participants risk selecting simulated humans that are only nearest to the AR cue due to the cue placement offset caused by latency in cue placement updates.

4.2 Experiment E2: Dropouts

Here we describe the second of the two experiments, evaluating the effects of network dropouts. See Section 4.1 for the description of the first experiment evaluating the effects of network latency.

4.2.1 Participants. The same participants as in Experiment E1 took part in this experiment (see Section 4.1.1). Participants completed E2 after E1. The inclusion of familiarization, the high number of task repetitions, and the simplicity of the search-and-selection task mitigate potential learning effects in task completion. The experiment took participants on average 30 minutes to complete.

4.2.2 Material. In this experiment, we followed a similar approach for dropouts as described for latency in Section 4.1.2. Here, we simulated network dropouts by controlling how often the AR tags that were shown on the AR HWD were updated. Dropouts were characterized by the AR tags freezing in place for the duration of the simulated network issue. Hence, the red diamond AR tags that were otherwise registered with and followed the movements of the simulated humans with the red armbands remained spatially in place. Once the dropout ended and the simulated network connection resumed, the red diamond AR tags immediately caught up with their registered positions over the heads of the humans with the red armbands.

We simulated three levels of network dropouts in this experiment. In order to select ecologically valid dropout ranges, we first looked at the literature for examples of severe dropout rates [37], leading to our decision to target a dropout level of 50% for the high dropout level. We also added an intermediate level with half the dropout

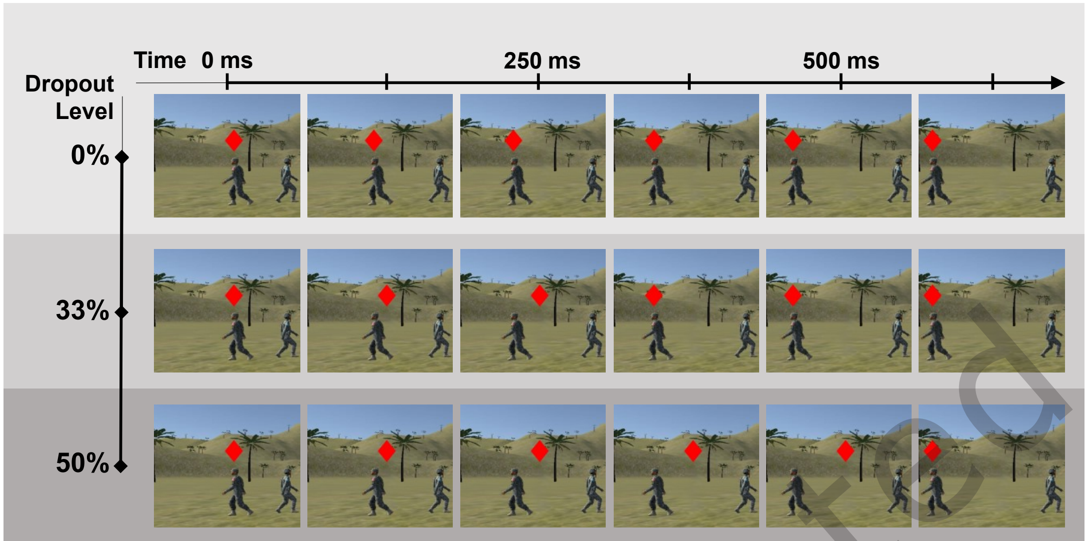


Fig. 8. Illustration of network dropouts in Experiment E2. Each row illustrates the positions of the red-team simulated human (left human) and their corresponding AR tag at different times. For illustration purposes, the position of the blue-team simulated human (right human) is fixed in these images, though they were also moving in the actual experiment. Network dropouts here are characterized by the position of the AR tag not being updated in some of the frames, i.e., the AR tags seem to be frozen in place during network dropouts even when the red-team simulated human is moving. The time scale refers to the duration of the dropouts, as discussed in Sec. 4.2.2.

interval as the high condition, or 25%. Together, we simulated the following levels of network dropouts (see also Figure 8 for an illustration):

- 0% dropout, no network dropouts.
- 33% dropout, 250 ms dropouts every 500 ms, meaning that 33% of the time the AR tags were not updated.
- 50% dropout, 500 ms dropouts every 500 ms, meaning that 50% of the time the AR tags were not updated.

We formally controlled AR network dropouts in this experiment with a timer in the rendering system as follows: (1) During time periods without dropouts, AR tags would move as nominal with the simulated humans. A timer starts for the amount of time specified between dropouts depending on the experimental condition. (2) The timer accumulates each frame by adding the amount of time between the current frame and the last frame. Each frame the timer checks if the total elapsed time is greater than or equal to the set time amount. (3) Once the timer expires, the position of AR tags cease to be updated, causing them to stay still as if the system providing information is experiencing a network dropout. (4) A new timer starts for the dropout time, as determined by the experimental condition. This timer behaves the same as in Step 2. (5) Once the timer from Step 4 expires, AR tag positions resume to be updated as before, returning to Step 1.

As also noted in Section 4.1.2, it is important to highlight that we simulated *network* dropouts but not *tracking* or *rendering* dropouts, i.e., while the positions of the AR tags in the environment around participants were not updated during network dropouts, the rendering on the AR HWD was still updated based on our AR head-tracking system. In other words, the AR tags remained spatially in-place in the environment when participants rotated their head in the experimental space.

Table 3. Experiment E2: Statistical test results for the dropout network factor and objective performance.

Measures	RM-ANOVA	Factors	df_G	df_E	F	p	η_p^2	Pairwise Comparisons
Elapsed Time	Two-way	Dropouts	2	38	9.50	$<$ **0.001**	0.33	All p**<0.05**: (33% Dropouts > 0% Dropouts), (50% Dropouts > 0% Dropouts)
		Behavior	3	57	3.73	**0.016**	0.16	All p**<0.05**: (Real-First > AR-First)
		Dropouts * Behavior	6	114	1.04	0.40	0.05	N/A
Type I Errors	Two-way	Dropouts	2	38	9.50	0.19	0.08	N/A
		Behavior	1.50	28.51	9.13	**0.002**	0.33	All p**<0.05**: (AR-Only > Real-Only), (AR-Only > Real-First)
		Dropouts * Behavior	1.43	28.33	0.96	0.46	0.48	N/A
Type II Errors	Two-way	Dropouts	2	38	0.46	0.64	0.02	N/A
		Behavior	1.26	23.95	10.65	**0.002**	0.34	All p**<0.05**: (AR-Only > Real-Only), (AR-Only > Real-First), (AR-Only > AR-First)
		Dropouts * Behavior	1.85	38.06	0.19	0.081	0.01	N/A

4.2.3 Study Design. To investigate trust in AR with respect to dropouts and different behaviors to perform the 360° search-and-selection task (see Section 3.2), we used a within-subjects design for our study with the same baseline and behaviors as in E1 in addition to 3 levels of Dropouts: high (50%), medium (33%), and low (0%). Behaviors were blocked and randomized, with the order of dropout conditions randomized within each behavior block. Each condition was tested twice, once while rotating clockwise and once in counterclockwise direction to avoid that our participants get entangled in the cables that were suspended from the ceiling.

4.2.4 Results. We analyzed the responses with repeated-measures analyses of variance (RM-ANOVAs) and Tukey multiple comparisons with Bonferroni correction at the 5% significance level. We confirmed 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 was not supported.

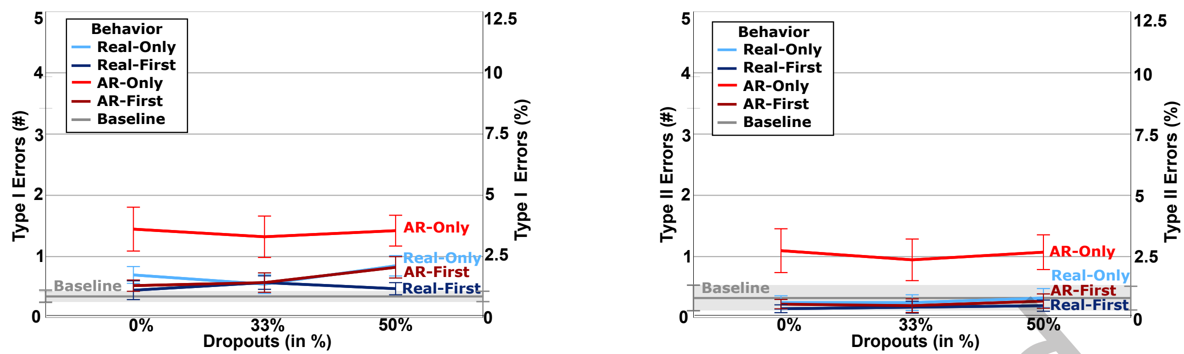
We found no significant difference between the clockwise and counterclockwise trials as well as the pre and post baselines, so we pooled the responses with respect to our analysis.

Objective Data. The descriptive statistics for the elapsed trial times are shown in Figures 9 and 10, and the statistical test results are shown in Table 3.

Our results show significant main effects for the dropout levels and behaviors on elapsed time, but no interaction effect between the two factors. Specifically, we found that elapsed times were significantly lowest (= best) for the 0% dropouts compared to the 33% and 50% dropout conditions. Further, we found that Real-First resulted in a significantly higher (= worse) elapsed time than the AR-First behavior.

Our results further show significant main effects for the behaviors on type I and II errors. Specifically, we found that type I errors were significantly higher for AR-Only than the Real-Only and Real-First behaviors. Moreover, we found that type II errors were significantly higher for AR-Only than the Real-Only, Real-First, and AR-First behaviors. The average error counts can be converted to error rates (20 pairs of simulated humans per trial and two trials per condition as discussed in Section 3.2) yielding error rates about 2.5% for each of Type I and Type II for the 50% dropout and AR-Only behavior condition.

Learning Effect Analysis. Learning effects were examined using linear mixed-effects models with trial index treated as a continuous within-participant predictor and participant modeled as a random intercept. For all analyses, the same model structure was used: `DV ~ Trial + (1 | PID)`. Models were fit using maximum



(a) Type I Errors

(b) Type II Errors

Fig. 9. Experiment E2: Results for (a) Type I errors and (b) Type II errors for the dropouts trials for the four behaviors and baseline condition. Lower is better. The error bars show the standard error.

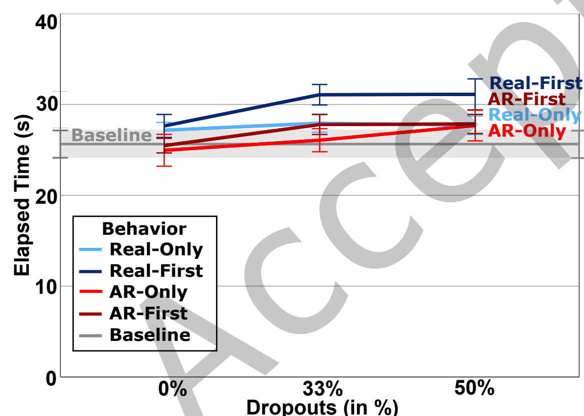


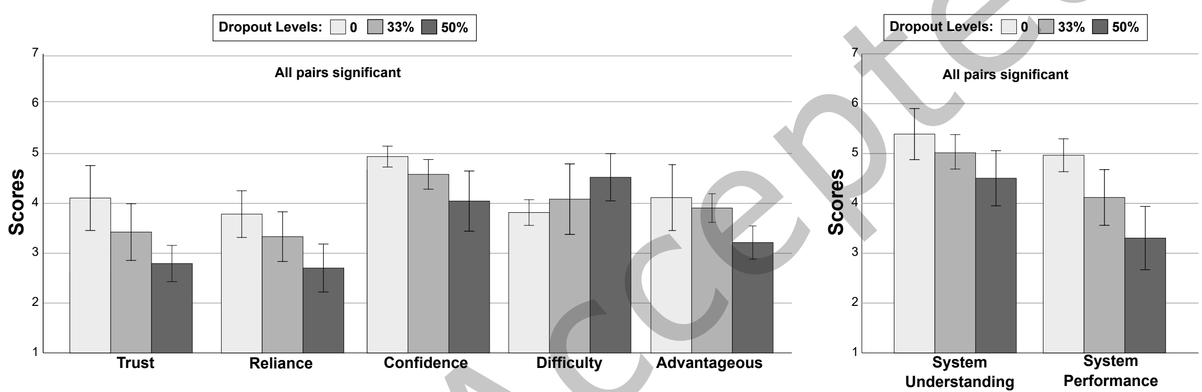
Fig. 10. Experiment E2: Results for elapsed time for the dropouts trials for the four behaviors and baseline condition. Lower is better. The error bars show the standard error.

likelihood estimation. The model predicting task completion time showed no significant effect of trial index ($\beta = -0.018$, SE $= 0.021$, $t = -0.86$, $p = .39$), indicating no systematic change in completion time across trials. The estimated random intercept variance for participant was $\sigma^2 = 19.44$. Trial index was not a significant predictor of Type I error rates ($\beta = 0.006$, SE $= 0.010$, $t = 0.61$, $p = .54$). The participant-level random intercept variance was $\sigma^2 = 0.31$. For Type II error rates, trial index did not significantly predict performance ($\beta = 0.004$, SE $= 0.009$, $t = 0.44$, $p = .66$), with participant-level random intercept variance $\sigma^2 = 0.29$.

Comparison to Baseline. To examine how each behavioral condition differed from baseline, we conducted within-participant comparisons using paired two-sided tests. For elapsed time, completion times differed significantly from baseline only for Real-First ($t(19) = 2.84$, $p = .010$, $d = 0.50$). No significant baseline differences were observed for AR-First ($p = .984$), AR-Only ($p = .941$), or Real-Only ($p = .962$). For Type I error, error rates were significantly higher than baseline for AR-First ($t(19) = 2.76$, $p = .012$, $d = 0.59$), AR-Only ($t(19) = 3.93$, $p < .001$, $d = 1.07$), and Real-Only ($t(19) = 2.67$, $p = .015$, $d = 0.67$). The difference between Real-First and

Table 4. Experiment E2: Statistical test results for the dropout network factor and subjective assessment.

Measures	**RM-ANOVA**	**Factors**	**df**$_G$	**df**$_E$	$\boldsymbol{F}$	$\boldsymbol{p}$	η_p^2	**Pairwise Comparisons**
TOAST System Understanding	One-way	Dropouts	1.10	21.00	25.14	<**0.001**	0.57	All pairs $\boldsymbol{p}$<**0.05**
TOAST System Performance	One-way	Dropouts	1.08	20.50	27.50	<**0.001**	0.59	All pairs $\boldsymbol{p}$<**0.05**
Trust	One-way	Dropouts	1.21	23.12	37.01	<**0.001**	0.66	All pairs $\boldsymbol{p}$<**0.05**
Reliance	One-way	Dropouts	1.05	19.89	17.59	<**0.001**	0.48	All pairs $\boldsymbol{p}$<**0.05**
Confidence	One-way	Dropouts	1.12	21.37	12.44	<**0.001**	0.39	All pairs $\boldsymbol{p}$<**0.05**
Difficulty	One-way	Dropouts	1.06	20.24	15.69	<**0.001**	0.45	All pairs $\boldsymbol{p}$<**0.05**
Advantageous	One-way	Dropouts	1.12	21.30	22.03	<**0.001**	0.54	All pairs $\boldsymbol{p}$<**0.05**



(a) Single-Item Questionnaires (b) TOAST Questionnaire

Fig. 11. Experiment E2: Subjective data results for the three levels of dropout with (a) Single-Item Questionnaires and (b) TOAST Questionnaire. Higher is better (except for Difficulty). All pairs were significant. The error bars show the standard error.

baseline was not statistically significant ($p = .242$). For Type II error, a significant difference from baseline was again observed only for AR-Only ($t(19) = 2.30$, $p = .033$, $d = 0.62$). Baseline comparisons for AR-First ($p = .606$), Real-First ($p = .490$), and Real-Only ($p = .828$) were not significant.

Subjective Data. The descriptive statistics for the subjective responses are shown in Figure 11, and the statistical test results are shown in Table 4.

Our results show significant main effects of the dropout levels on both subscales of the TOAST questionnaire as well as our single-item questionnaires for Trust, Reliance, Confidence, Difficulty, and Advantageous. All pairwise comparisons were significant. Overall, participants exhibited significantly less trust, reliance, and confidence in the AR system the higher the network dropouts were. They further felt that the task was more difficult and the AR system less advantageous the higher the network dropouts were.

4.2.5 Discussion.

Network Dropouts. The results of Experiment E2 indicate that the network dropout levels of the AR tags had a significant effect on participants' task performance for our tested search-and-selection task. Adding dropouts caused participants to take more time to complete the task (see Figure 10), but it is interesting that it did not cause them to make more errors (see Figure 9), which stands in stark contrast to the results for network latency discussed in Section 4.1.5.

A potential explanation for these findings is that increased dropouts meant that participants needed more time to determine which human the AR tag was registered to. The fact that participants did not make more errors in high-dropout conditions supports this claim, i.e., when participants detected a dropout occurred (which was obvious because the simulated humans in our study were consistently moving), they waited for the dropout to resolve before selecting a human. However, it is further interesting to note that we do not see a linear increase in elapsed time due to network dropouts but rather a plateau effect for higher dropouts, suggesting that participants were able to adapt to the dropouts and complete the task with comparatively high performance even for high dropouts.

Based on the subjective questionnaire data, participants trusted the AR system significantly less as the network dropout levels increased. Further, they also indicated that dropouts significantly affected their reliance on AR, how difficult they perceived the task, their confidence in their task performance, and how advantageous they perceived the AR system for the task.

Behaviors. Behavior had a significant main effect on participants' task completion time and errors. Specifically, Real-First was slower than AR-First, and AR-Only resulted in more errors than other behaviors. In contrast to the results for latency in Experiment E1, it is interesting to see that dropouts did not introduce additional errors for the AR-Only behavior beyond its intrinsic base error rate.

However, it can be seen that these errors were greatly reduced when participants used the AR-First or Real-First behaviors, i.e., when participants integrated cues from both information channels (AR tags vs. armbands). Similar to our results for latency in Experiment E1, the AR-First behavior showed significantly faster elapsed times compared to Real-First. Hence, as discussed in Section 4.1.5, if faced with the decision between one of these behaviors, it seems that the AR-First behavior is generally preferable over the Real-First behavior, and both are preferable over the AR-Only behavior.

4.3 General Discussion

In this section, we summarize the main results from our two experiments and provide a general discussion.

4.3.1 Higher Task Performance for Lower Latency and Dropouts. The appearance of both latency or dropouts caused participants to take longer to complete the tasks. While the high latency condition elicited longer task completion times compared to the intermediate latency condition, the high dropout condition did not generate significantly higher elapsed times compared to the intermediate dropout condition.

The result that elapsed time increased as latency increased may indicate that participants do not adopt a fixed-time strategy to overcome the registration error that latency introduces, as has been explained in prior work [18]. When the distance between the AR cue and the simulated human increases, then it may become more difficult to transfer attention back and forth between the cue and simulated human, increasing the time to complete the task. Since higher latency increases the distance between the moving simulated human and the AR cue, it could then be expected that task completion time would increase, as well.

This explanation also addresses why task completion time with dropouts does not increase with additional dropouts. Increasing dropout time does not uniformly increase the distance between the AR cue and the simulated human target, as the distance varies depending on how long it has been since the dropout began. Furthermore, during the time that dropouts are not occurring, the AR cue is perfectly registered to the target. Therefore, for a large portion of the interaction the distance between the AR cue and the simulated human target does not increase with dropout time. Even though the search-and-selection task is the same as in the latency conditions, we may not expect task completion time to increase with dropouts the same way as with increases in latency for this reason.

For both type I and II errors, we found that latency level had a significant main effect while dropout level did not (see Tables 1 and 3). One explanation for this difference is similar to the explanation for the differences between the two factors with regards to task completion time. Increasing latency creates a greater error in the registration of the AR cue, making the task more difficult, as discussed above and in Section 4.1.5. On the other hand, when participants encountered dropouts the AR cues were correctly registered to the simulated human for part of the time, allowing participants to avoid errors by observing the position of the cue during this time, as discussed above and in Section 4.2.5.

Thus, one explanation for the differences in the effects of latency and dropouts on participants' task performance is due to differences in how latency and dropouts affect the distance between the cue and the target simulated human over the duration of the task, with latency consistently creating larger distance with larger latency and dropouts creating no distance at all for part of the interaction.

4.3.2 Higher Trust, Reliance, Confidence, and Perceived Advantages of the AR System for Lower Network Latency and Dropouts. TOAST ratings on both dimensions, subjective responses to our trust, reliance, confidence, and advantageous questions all decreased, i.e. declined, as latency and dropout duration increased. This is expected, as problems affecting registration are usually obvious to the user and directly impact user experience. It is notable that all pairs of dropout level demonstrated a significant effect on these subjective ratings, in contrast with objective measures where there were only significant differences between (0%, 25%) dropout and (0%, 50%) dropout. One explanation for this discrepancy is a mismatch between an awareness of additional system degradation, i.e. worsened dropout duration, versus the robustness of user strategy to the degradation, e.g. using the periods where there were no dropouts to successfully complete the task. Users perceive a decrease in trustworthiness, desire to rely, confidence in their own performance, and advantageousness of the system, as well as an increase in task difficulty, as dropout duration increases despite using the system to equal success between the two dropout conditions. These subjective factors may lead to changes in reliance choices in the real-world, where we cannot always enforce behaviors on end users. For example, even if an interface is designed to be objectively performant in poor network conditions, users may still choose to disuse the system due to the perception that poor network conditions have compromised the usefulness of the system. It will be necessary to also design networked AR systems to support both robust objective performance and robust subjective evaluations during network difficulties.

4.3.3 Mismatches Between Perceived Difficulty and Task Performance. For perceived difficulty, latency level did not show a significant effect while dropouts did. This contrasts with our objective results, where task performance decreased as latency increased but did not significantly vary between the low and high dropout condition. It would be expected that subjective results mirror objective performance measures; subjective ratings decrease as latency increases and subjective ratings stay the same for low and high dropout conditions. Instead, we see the opposite effect: difficulty ratings do not change by latency condition and all ratings are significantly affected by dropout level. In prior work, users incorrectly perceived latency to be minimally impactful on their task performance, despite objective results showing a clear detrimental effect of latency on performance [18]. It is possible that participants found the task of compensating for latency offsets to be easy to understand and execute, despite its clear effects on objective performance. It is well known that users do not always properly understand the relative effects of user interface factors on performance [28]. In this case, participants' perception of the difficulty that latency introduces may rely on heuristics and assumptions that are not aligned with the actual nature of the task at hand. It is also notable that latency introduces a greater possibility for error compared to dropouts. In the latency conditions, the AR cues are almost never correctly placed, leading to predictably high error rates when using the AR-Only behavior. We do not observe this interaction for dropouts, indicating that the nature of the effects introduced by latency is more prone to causing errors.

With dropouts, we see another mismatch between perceived difficulty and objective performance, but in reverse: the objective performance did not change between the medium and high dropout conditions, i.e. the 33% and 50% dropout conditions, although difficulty ratings significantly increased between them. It is possible that the same explanation for the mismatch between perceived difficulty and performance in the latency conditions applies to dropout, that participant perception the impact of dropouts on task difficulty is not calibrated with factors that make the task harder to complete.

Taken together, these results demonstrate that even a seemingly simple task such as search-and-selection can trigger complicated relationships between subjective experience and human performance when subjected to relatively straightforward network degradations like latency and dropouts. It is important to manage network performance to mitigate the risk that poor network performance may contribute to disconnects between a user's subjective evaluation of their task and the reality of the work that is being done, potentially impacting decision making based on these factors.

4.3.4 Highest Trust and Reliance on AR Does Not Necessarily Result in Highest Task Performance, But Verification Behaviors Mitigate Errors. Our evaluations of network factors in Experiments E1 and E2 indicate that participants made significantly more errors with the AR-Only behavior compared to the other behaviors in the context of a dynamic task with cluttered targets. Ostensibly, this is due to the fact that the AR-Only behavior prevents users from using information from the armband and only depend on the AR cues. So AR cues affected by latency and dropout errors resulted in higher errors. Thus in real-world situations where increased latency and connectivity interruptions are expected, it is important that users do not demonstrate complete reliance on AR cues.

In the Real-First and AR-First behaviors, participants were able to compare AR cue information with the environment and use both sources of information in the search-and-selection. In both experiments, AR-First behaviors resulted in faster task completion times than Real-First behaviors without increasing type I or II errors. This is interesting because both behaviors integrate information from the same sources, AR cue and environment. One explanation for this is the difference in the saliency between the AR cues and the simulated humans. Despite the registration errors caused by latency and dropouts, the movement of the AR cues and simulated humans provide information about each other. The movement of the AR cues can tell the user how fast the simulated humans move and what their paths look like, and vice versa. Yet the AR cues are more salient than the simulated humans wearing uniforms, potentially making this movement information easier to understand and easier to convert into cues on simulated human positioning than using information on simulated human movement to understand the movement of the AR cues. This mirrors prior work on "naive realism" where simpler, functional displays outperform realistic ones by displaying only the relevant information for the task [28]. In this case, the walking simulated humans contain visual information not just on their armband color, but also on their uniform pattern, the animation of their appendages, and what way they are facing, all of which is not useful to the user whose next task will be to locate the AR cue. Thus the AR-First behavior results in faster task performance times due to the simple AR cues communicating target movement cues better than the more realistic simulated humans. Future work may further investigate the differences between the Real-First and AR-First behaviors to elucidate the mechanism behind this performance disparity.

4.3.5 Limitations and Future Work. In our experiments we evaluated the effects of two network factors with three levels each. Due to the large number of conditions and complexity of our experiments, we were not able to compare more than three levels of latency and dropouts. We decided on the levels we tested based on practical considerations with Internet/network connections, 4G/5G cellphone connections, the literature, and our own experience, but we acknowledge that we did not include further extremes and in-between values. We make no claims about network factors beyond the values we tested in our experiments. Interested researchers may evaluate these aspects in future work. For instance, even higher network performances may be interesting to evaluate in the context of edge processing or peer-to-peer AR system networks, and the effects of even lower network performances may be interesting in the context of cloud processing or contested AR network environments. Further, different methods may be used to compensate for and reduce the effects of network latency or dropouts, such as predictions based on extrapolated linear or nonlinear AR tag movements. It would be interesting to evaluate the tradeoffs with such compensation methods related to AR users' task performance and subjective assessments of the system.

In our procedure, we trained and prompted participants to ensure they understood the behaviors they were to implement during the experiment. Future work could employ techniques such as gaze tracking to independently verify that participants were employing the behaviors as instructed.

There are many factors that can potentially impact visual search-and-selection performance including gender [14, 33] and prior experience with AR devices. These factors may have an influence on performance with one behavior over another, as well as the actual selection of behaviors in the real world. Though we did not find a specific reason to expect gender or other factors to affect the degree to which latency and dropouts could affect user performance and trust, future work can investigate the effects of these and other factors on trust behaviors through techniques such as the use of balanced gender samples and the employment of surveys accounting for experience with AR devices.

The search-and-selection task we employed was very easy in that the baseline task is easy to complete unaided. It would be interesting to see how task performance and attitudes towards an AR system with network errors would change in a scenario where completion of the task is very difficult without the AR system. Such tasks could include identifying targets in brownout, whiteout, or blackout conditions; targets in a crowd; and targets whose movement makes identifying visual features becomes difficult.

We administered the subjective self-report questionnaires at the end of all trials for all conditions together, as outlined in Section 3.3.3. This has the advantage of making it easier for participants to differentially rate conditions relative to each other but not going so far as to force a ranking, as well as shortening the overall length of the experiment. This method has the limitation that their ratings are prefaced by the sum of the experiences and may not fully reflect their opinions at the moment of completing each trial. This is contrasted with the objective results, which by their nature reflect performance as it happens. The mismatch between subjective and objective results discussed earlier may change if questionnaires were administered directly after each condition.

5 Conclusion

In this paper we presented the results of two experiments we conducted to evaluate the effects of network errors on AR users' objective performance and subjective assessments of the technology in the scope of connected AR HWD systems in which spatial task-relevant cues are provided over a network connection. Our results highlight the importance of considering network factors in connected AR systems, their effects on objective performance, subjective perception, trust, and reliance, and implications they have on the choice and effectiveness of different behaviors when completing real-world tasks supported by AR information. The results provide insights and implications for the design and development of connected AR systems and applications in a wide range of task contexts.

Acknowledgments

This material includes work supported by the US Air Force Research Laboratory (AFRL) under an SBIR effort (contract number FA864922P1038) and was supported in part by the National Science Foundation under award numbers 2235066 and 1800961 (Dr. Ephraim P. Glinert, IIS), the Office of Naval Research under award numbers

N00014-21-1-2578 and N00014-21-1-2882 (Dr. Peter Squire, Code 34), and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

References

[1] NGMN Alliance. 5g white paper. *Next generation mobile networks, white paper*, 1, 2015.

[2] ATIS. ATIS Telecom Glossary, 2007.

[3] Tristan Braud, Farshid Hassani Bijarbooneh, Dimitris Chatzopoulos, and Pan Hui. Future Networking Challenges: The Case of Mobile Augmented Reality. In *International Conference on Distributed Computing Systems (ICDCS)*, pages 1796–1807, 2017.

[4] Austin Erickson, Nahal Norouzi, Kangsoo Kim, Ryan Schubert, Jonathan Jules, Joseph J LaViola Jr, Gerd Bruder, and Gregory F Welch. Sharing gaze rays for visual target identification tasks in collaborative augmented reality. *Journal on Multimodal User Interfaces*, 14(4):353–371, 2020.

[5] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007.

[6] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal Technologies*, 1:208–217, 1997.

[7] Paolo Ferrari, Emiliano Sisinni, Dennis Brandão, and Murilo Rocha. Evaluation of communication latency in industrial iot applications. In *IEEE International Workshop on Measurement and Networking (M&N)*, pages 1–6, 2017.

[8] Tharaka Hewa, Gürkan Gür, Anshuman Kalla, Mika Ylianttila, An Bracken, and Madhusanka Liyanage. The role of blockchain in 6G: Challenges, opportunities and research directions. *IEEE 6G Wireless Summit (6G SUMMIT)*, pages 1–5, 2020.

[9] Richard L. Holloway. Registration error analysis for augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):413–432, 1997.

[10] Tongyi Huang, Wu Yang, Jun Wu, Jin Ma, Xiaofei Zhang, and Daoyin Zhang. A survey on green 6G network: Architecture and technologies. *IEEE Access*, 7:175758–175768, 2019.

[11] Dongsik Jo and Gerard Jounghyun Kim. AR enabled IoT for a smart and interactive environment: A survey and future directions. *Sensors*, 19(19), 2019.

[12] Kangsoo Kim, Mark Billinghurst, Gerd Bruder, Henry Been-Lirn Duh, and Gregory F. Welch. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2947–2962, 2018.

[13] Kangsoo Kim, Nahal Norouzi, Dongsik Jo, Gerd Bruder, and Gregory F Welch. The augmented reality internet of things: Opportunities of embodied interactions in transreality. In *Springer Handbook of Augmented Reality*, pages 797–829. 2023.

[14] Jillian E. Lauer, Eukyung Yhang, and Stella F. Lourenco. The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin*, 145(6):537–565, 2019.

[15] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[16] Siyi Liao, Jun Wu, Jianhua Li, and Kostromitin Konstantin. Information-centric massive IoT-based ubiquitous connected VR/AR in 6G: A proposed caching consensus approach. *IEEE Internet of Things Journal*, 8(7):5172–5184, 2020.

[17] Peter Lincoln, Alex Blate, Montek Singh, Turner Whitted, Andrei State, Anselmo Lastra, and Henry Fuchs. From Motion to Photons in 80 Microseconds: Towards Minimal Latency for Virtual and Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1367–1376, 2016.

[18] Mark A Livingston and Zhuming Ai. The effect of registration error on tracking distant augmented objects. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 77–86, 2008.

[19] Roger C. Mayer, James H. Davis, and F. David Schoorman. An Integrative Model Of Organizational Trust. *Academy of Management Review*, 20(3):709–734, July 1995. Publisher: Academy of Management.

[20] Sebastian Merker, Stefan Pastel, Dan Bürger, Alexander Schwadtke, and Kerstin Witte. Measurement Accuracy of the HTC VIVE Tracker 3.0 Compared to Vicon System for Generating Valid Positional Feedback in Virtual Reality. *Sensors*, 23(17), 2023.

[21] James Louis Merlo. *Effect of reliability on cue effectiveness and display signaling*. University of Illinois at Urbana-Champaign, 1999.

[22] Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, 8(1):47–63, 1998.

[23] Diederick C. Niehorster, Li Li, and Markus Lappe. The Accuracy and Precision of Position and Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research. *i-Perception*, 8(3), 2017.

[24] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[25] Andriy Pavlovych and Wolfgang Stuerzlinger. Target following performance in the presence of latency, jitter, and signal dropouts. In *Graphics Interface (GI))*, pages 33–40. 2011.

[26] Michelle L Rusch, Mark C Schall Jr, Patrick Gavin, John D Lee, Jeffrey D Dawson, Shaun Vecera, and Matthew Rizzo. Directing driver attention with augmented reality cues. *Transportation research part F: traffic psychology and behaviour*, 16:127–137, 2013.

[27] Ryan Schubert, Greg Welch, Salam Daher, and Andrew Raij. HuSIS: a dedicated space for studying human interactions. *IEEE Computer Graphics and Applications*, 36(6):26–36, 2016.

[28] Harvey S. Smallman and Mark St. John. Naive Realism: Misplaced Faith in Realistic Displays. *Ergonomics in Design*, 13(3):6–13, 2005.

[29] Robert D Sorkin. Why are people turning off our alarms? *The Journal of the Acoustical Society of America*, 84(3):1107–1108, 1988.

[30] Kate A. Spitzley and Andrew R. Karduna. Feasibility of using a fully immersive virtual reality system for kinematic data collection. *Journal of Biomechanics*, 87:172–176, 2019.

[31] Youngjung Suh, Youngmin Park, Hyoseok Yoon, Yoonje Chang, and Woontack Woo. Context-Aware Mobile AR System for Personalization, Selective Sharing, and Interaction of Contents in Ubiquitous Computing Environments. In *International Conference on Human-Computer Interaction*, pages 966–974, 2007.

[32] Nicholas Toothman and Michael Neff. The impact of avatar tracking errors on user experience in vr. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 756–766. IEEE, 2019.

[33] Wenheng Wang, Mingyu Zhang, Zhide Wang, and Qing Fan. The Impact of Spatial Dimensions, Location, Luminance, and Gender Differences on Visual Search Efficiency in Three-Dimensional Space. *Buildings*, 15(5):656, 2025.

[34] Amelia C Warden, Christopher D Wickens, Domenick Mifsud, Shannon Ourada, Benjamin A Clegg, and Francisco R Ortega. Visual search in augmented reality: Effect of target cue type and location. In *Human Factors and Ergonomics Society Annual Meeting*, volume 66, pages 373–377, 2022.

[35] Greg Welch, Gerd Bruder, Peter Squire, and Ryan Schubert. Anticipating Widespread Augmented Reality: Insights from the 2018 AR Visioning Workshop. Technical report, University of Central Florida and Office of Naval Research, August 6 2019.

[36] Heather M Wojton, Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan. Initial validation of the trust of automated systems test (TOAST). *The Journal of Social Psychology*, 160(6):735–750, 2020.

[37] Xuesu Xiao and Shuayb Zarar. Packet loss concealment with recurrent neural networks for wireless inertial pose tracking. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 25–29, Las Vegas, NV, March 2018. IEEE.

[38] Michelle Yeh. *Attention and trust biases in the design of augmented reality displays*. University of Illinois at Urbana-Champaign, 2000.

[39] Michelle Yeh and Christopher D Wickens. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3):355–365, 2001.

[40] Wenxiao Zhang, Bo Han, and Pan Hui. On the networking challenges of mobile augmented reality. In *Workshop on Virtual Reality and Augmented Reality Network*, pages 24–29, 2017.

[41] Lin Zhu, Suryadipta Majumdar, and Chinwe Ekenna. An invisible warfare with the internet of battlefield things: A literature review. *Human Behavior and Emerging Technologies*, 3(2):255–260, 2021.