Remote research on locomotion interfaces for virtual reality: Replication of a lab-based study on teleporting interfaces

Jonathan W. Kelly, Melynda Hoover, Taylor A. Doty, Alex Renner,
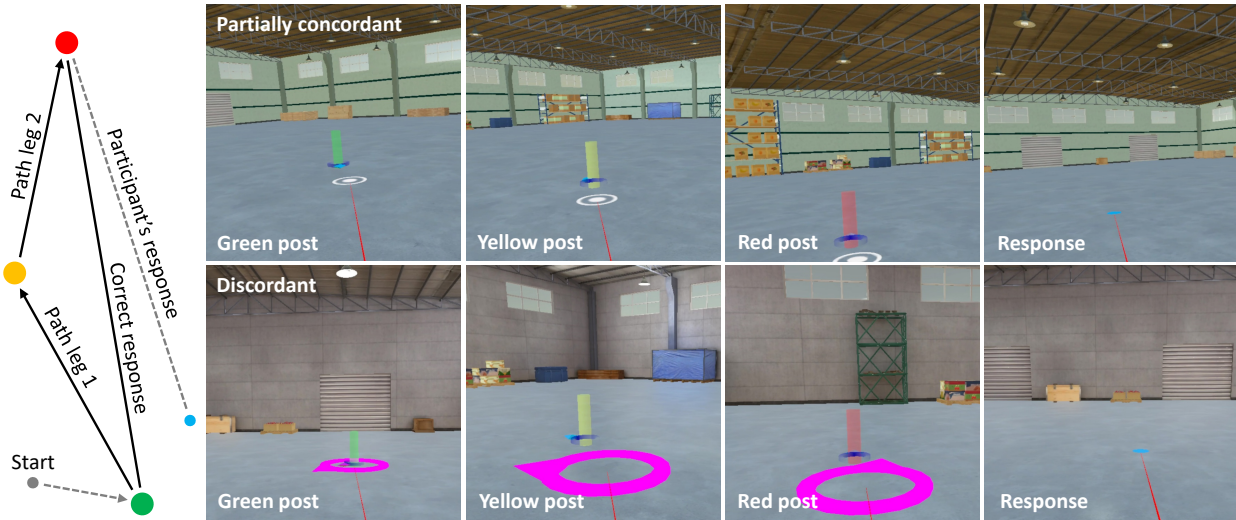Moriah Zimmerman, Kimberly Knuth, Lucia A. Cherep, and Stephen B. Gilbert

Fig. 1. The diagram on the left depicts the triangle completion task. The participant traveled sequentially to three locations marked by vertical posts (green, then yellow, then red) before attempting to point to the remembered location of the green post. Screenshots show perspective views of the three posts marking the triangle as well as the response. The top row shows the partially concordant teleporting interface, where the participant teleports to translate but rotates using the body. The top row also depicts the large virtual environment. The bottom row shows the discordant teleporting interface, where the participant teleports to translate and rotate. The bottom row also depicts the small virtual environment.

Abstract—The wide availability of consumer-oriented virtual reality (VR) equipment has enabled researchers to recruit existing VR owners to participate remotely using their own equipment. Yet, there are many differences between lab environments and home environments, as well as differences between participant samples recruited for lab studies and remote studies. This paper replicates a lab-based experiment on VR locomotion interfaces using a remote sample. Participants completed a triangle-completion task (travel two path legs, then point to the path origin) using their own VR equipment in a remote, unsupervised setting. Locomotion was accomplished using two versions of the teleporting interface varying in availability of rotational self-motion cues. The size of the traveled path and the size of the surrounding virtual environment were also manipulated. Results from remote participants largely mirrored lab results, with overall better performance when rotational self-motion cues were available. Some differences also occurred, including a tendency for remote participants to rely less on nearby landmarks, perhaps due to increased competence with using the teleporting interface to update self-location. This replication study provides insight for VR researchers on aspects of lab studies that may or may not replicate remotely.

Index Terms—Navigation, Spatial cognition, Virtual reality, Teleporting, Online data collection

---◆---

1 INTRODUCTION

The rising popularity of consumer-oriented virtual reality (VR) equipment has created a new landscape for VR research, which can now be conducted by recruiting owners of head-mounted displays (HMDs) who participate remotely using their own equipment. There are many advantages to remote VR research. For example, remote research provides greater safety to researchers and participants during the COVID-19 pan-

demic. Remote research also provides opportunities to study different populations, such as individuals with considerable VR experience.

However, there are numerous differences between carefully controlled lab environments and home environments of VR owners. For example, lab environments can provide open walking space to enable participants to freely explore the virtual environment (VE) on foot, but such spaces are uncommon in the home. Labs can also be kept quiet and consistent across participants, whereas homes often contain distraction from roommates, family, pets, and technology, and there will be wide variation in home environments across participants.

There may also be important differences between participant samples recruited for lab and remote studies [10, 34]. For example, remote participants are likely to be older and more diverse than those recruited through the typical university community. Also, HMD owners spend more time playing video games than non-owners, and women are underrepresented among HMD owners compared to non-owners.

• The authors are with Iowa State University, Ames, IA, 50011. E-mail:
{jonkelly, mthoover, tdoty, arenner, moriahz, gilbert} @iastate.edu,
lacherep@arizona.edu, kimberly.n.knuth@gmail.com.

Given the numerous differences between lab research and remote research, a cautious approach is warranted to determine whether results generalize between laboratory VR studies and remote VR studies. Science builds on past research, and replication of in-person research when using a remote sample is critical to the interpretation and acceptance of remotely conducted VR research studies.

This paper presents an experiment using a remote sample to replicate a lab-based experiment on locomotion interfaces [11]. The remote study evaluated performance on a navigation test called triangle completion (see Figure 1), in which participants traveled two legs of an outbound path before attempting to point to the path origin. Travel along the outbound path was accomplished using two forms of the teleporting interface. To teleport, the user selects a place in the VE and is then instantly transported to that location. Teleportation is commonly used in VR to avoid collisions with real objects and to reduce sickness. Teleportation can be disorienting due to the lack of self-motion cues [4], but nearby environmental cues such as landmarks and boundaries mitigate disorientation [6, 7, 12]. The replication study included several manipulations that can impact disorientation, including the size of the triangle and the size of the surrounding VE, which collectively determine the proximity of environmental cues to the task space. The results are used to evaluate whether primary findings from a lab-based study on locomotion in VR replicate under remote unsupervised testing conditions.

2 RELATED WORK

This section reviews past VR research involving remote participants, followed by a description of how teleporting interfaces impact navigation.

2.1 Remote VR Research

There is relatively little published research in which VR studies have been conducted remotely. This is likely to change quickly as a result of the COVID-19 pandemic, during which VR researchers were forced to pivot from traditional in-person lab-based research to remote research. A CHI 2021 workshop entitled Remote XR Studies included 13 presentations on the topic. A recent commentary [29] spurred by the pandemic identified three solutions for conducting VR research remotely: 1) inter-lab agreements in which lab personnel use lab equipment to collaboratively complete studies conducted by another lab, 2) distribution of VR equipment to participants in the researcher's community, and 3) recruitment of existing owners of VR equipment. Although each solution has several pros and cons, the third solution is the most popular among recent remote VR studies. The current research project followed this trend by recruiting existing owners of VR equipment, and this section summarizes VR research that has taken this approach.

A major concern when conducting remote VR research is how to interpret findings in the context of the existing literature, which is largely based on lab-based research. This is ultimately a question about generalizability of findings between lab-based and remote research. A few studies have explicitly attempted to replicate lab-based (and sometimes real-world) research findings using a remote sample of participants who own their own VR equipment. In one study [34], lab and remote samples first attempted to memorize locations of 3D points within a VE and, after exiting the VE, performed memory-based distance judgments about inter-point separations. Memorization time was faster for remote participants compared to lab participants, which could reflect differences in spatial ability or differences in familiarity with the controllers and equipment used in this stage of the task. Subsequent memory-based judgments were comparable between the remote and lab samples, and both samples were affected by manipulations involving alignment between the studied perspective and the axis of separation between the judged points.

In another study [26], participants completed a 3D Fitts's law pointing task using VRChat, a social VR platform. Participants pointed to objects varying in distance and size while supervised by a researcher, who was present in the VRChat environment and represented by an avatar. The results generally followed past findings on Fitts's law, such

as the relationship between pointing time and target size/distance, although details of specific model fits for the 3D pointing did not perfectly replicate past lab-based VR research [2].

Another replication study from the same paper [26] investigated collaborative work on a tabletop task by pairs of participants in a remote VR study, also using VRChat. This study examined time spent working together as well as specific behaviors such as gestures and viewing positions (e.g., two participants next to each other, two participants on opposite sides of the task table, etc.) as a function of task type. The results generally matched those of the original lab-based real world study upon which the study was based [30]. However, there were also subtle differences, such as proportion of time spent in certain viewing position arrangements. The authors attributed these subtle differences to the many differences between their virtual collaborative task and the real-world collaborative task that it was based on.

Some remote VR replication attempts have been less successful. For example, one paper [15] attempted to replicate three research findings in separate remote VR studies, but two of the three attempts were unsuccessful. One of the studies attempted to replicate a real-world finding that a nature walk increases positive feelings compared to an urban walk [8, 31], but no difference was found between environments in the remote VR version of the study. Another study attempted to replicate the finding that participants negotiate more aggressively when represented by a taller self-avatar, a study that was originally conducted with a lab-based sample [32]. However, participant height did not affect negotiation in the remote VR version [15] (but see [9] for a partial replication of this finding using a remote sample). Of the three attempts, the only successful replication was the finding that a crowd of avatars attracts participant attention in a manner similar to that found with a crowd of real people [16]. These results indicate that replication of lab results in a remote setting cannot be assumed, and that further work is needed to determine the key ingredients of a successful remote replication.

Other remote VR research studies have produced novel results on topics such as the influence of instructor avatars on learning [23] and the impact of a self-avatar on presence and embodiment [28]. However, such projects were not replications of lab-based research, and therefore are beyond the scope of the current review.

One notable difference between remote VR studies and lab-based studies is the shift in participant demographics. A survey comparing VR headset owners to non-owners [10] reported that women are underrepresented among VR owners. Underrepresentation of women has been documented even in lab-based VR research [22], and this problem is likely to be compounded in remote VR research that recruits VR owners. In one recent and relatively large (n=226) remote VR study [17], women comprised only 12% of the final sample. Another potentially important sample difference is that VR headset owners, on average, spend more time playing video games than do non-owners [10]. Demographic differences such as these are particularly relevant to VR research due to 1) the inherently spatial nature of VR and 2) the documented associations between spatial ability and participant characteristics such as gender and video game experience [6, 10].

To summarize, there is a paucity of research in which known findings from lab-based VR studies or real world studies have been replicated in a remote VR setting. Yet, this approach is absolutely necessary if new results from remote VR studies are to be integrated into the existing knowledge base formed by lab-based VR research. The few studies to attempt replication of lab-based results in a remote VR setting have produced mixed conclusions, with some findings replicating and others not. As studies continue to draw on lab vs. remote samples, it will be important to note characteristics of those sample populations that may affect results.

2.2 Spatial Updating and Teleporting Interfaces

Spatial updating is the process of updating self-location and self-orientation during locomotion. In other words, spatial updating involves keeping track of one's changing place relative to the environment. This process is central to basic navigational tasks such as avoiding obstacles and progressing toward goals.

Triangle completion is a common test of spatial updating. In a typical triangle completion task, the navigator travels two legs of a pre-defined outbound path before attempting to return to or point to the path origin (i.e., to complete the triangle). If the navigator is able to accurately update self-position while traversing the outbound path, then pointing to the path origin will be relatively trivial and errors will be small. However, if spatial updating is disrupted (e.g., by degradation of self-motion cues experienced along the outbound path), then the navigator will have greater uncertainty about their position within the environment and pointing errors will be consequently larger.

Spatial updating relies primarily on two inputs: path integration and piloting. Path integration is the process of integrating self-motion cues over time. Cues to path integration include vestibular input from the inner ear, proprioceptive input from limbs and joints, and visual and auditory self-motion signals (optic flow and acoustic flow, respectively). Removal of self-motion cues, as when passively guided in a wheelchair, has a negative effect on triangle completion performance [1].

Piloting relies on recalling landmark locations and identifying self-position relative to remembered landmarks. Using the triangle completion task as an example, path integration can be used to update the remembered location of the path origin while traversing the outbound path, and piloting can be used to recall the location of the path origin on the basis of environmental cues. Input from path integration and piloting are commonly combined to achieve a single estimate of self-location [5, 20, 27, 33].

Most VEs have ample landmarks to enable piloting, but the self-motion cues available for path integration depend heavily on the locomotion interface. For example, forward movement with a purely joystick-based interface preserves visual self-motion cues (primarily optic flow), but lacks the vestibular, proprioceptive, and kinesthetic cues normally associated with real-world locomotion. Even a treadmill interface that preserves forward leg motions still lacks vestibular cues associated with translational acceleration. Teleporting interfaces — the emphasis of the current project — involve discrete jumps in position and/or orientation, and therefore lack all cues normally associated with self-motion.

The concordance framework [7] was proposed to describe the extent to which movement of the body is consistent (i.e., concordant) with movement through the VE. This framework emphasizes the importance of body-based self-motion cues to the process of spatial updating [13, 21, 24, 25]. For example, in one form of the teleporting interface, the user selects a position and orientation in the VE by positioning and orienting a reticle and is then instantly transported to that position and orientation. This interface is referred to as **discordant teleporting** because movement through the VE is in complete disagreement with movement of the body. That is, the body is completely stationary despite changes in position and orientation of the viewpoint within the VE. The discordant teleporting interface as implemented in the current study is shown in the bottom row of Figure 1.

In another version of the teleporting interface, the user teleports to change their position in the VE but rotates their body to rotate in the VE. This interface is referred to as **partially concordant teleporting** because rotations in the VE are consistent with movement of the user's body but translations in the VE are in disagreement with the user's stationary body. In other words, some movements are concordant and others are discordant, hence the term partially concordant. The partially concordant teleporting interface as implemented in the current study is shown in the top row of Figure 1.

Both forms of teleporting, discordant and partially concordant, cause disruption of path integration that cannot be fully compensated for by reliance on piloting [4, 7, 11, 12]. However, discordant teleporting leads to greater disorientation than does partially concordant teleporting [7, 11, 12], due to the lack of rotational self-motion cues with discordant teleporting. This has downstream effects on the user, including lower accuracy in their cognitive map of the VE that they traversed [14].

Connecting this topic back to research on remote VR (Section 2.1), one study [6] related triangle completion performance when teleporting to participant characteristics such as gender and spatial ability— characteristics that are expected to be biased by recruitment of a remote participant sample compared to a lab sample [10]. Compared to participants who performed poorly on the navigation task, those who performed well were predominantly male, played more hours of video games, and performed better on tests of spatial ability. This is particularly worrisome given that VR headset owners, compared to lab-based participants, play more hours of video games and are more likely to be male.

3 EXPERIMENT OVERVIEW

Using a remote participant sample to replicate lab-based VR findings on locomotion interfaces will provide a crucial link between new results collected remotely and existing findings from lab-based studies. Spatial updating, measured using the triangle completion task, is a core navigational competency and therefore provides an ideal foundation for comparison between remote and lab-based research.

There are several reasons why spatial updating performance when using teleporting interfaces might differ across remote and lab-based research. VR headset owners typically have considerable experience with several locomotion interfaces, including teleportation, whereas lab-based studies usually include fewer VR experts and often include many VR novices (e.g., in a recent study with 182 undergraduate student participants, only 15 or 8% reported having any prior VR experience [6]). Furthermore, VR owners are familiar with the myriad controller buttons that sometimes overwhelm VR novices who participate in lab-based research. As described in Section 2.1, VR owners may also possess characteristics (e.g., gaming experience) that predispose them to succeed in spatial tasks. Collectively, these differences suggest that VR owners may perform better than the typical lab participants at locomotion tasks in VR. Alternatively, the unsupervised nature of remote research may lead to attention lapses, interruptions, or carelessness, thereby adding noise to the data and obfuscating differences between experimental conditions that would otherwise emerge in a lab setting. Furthermore, performance by remote participants could be impaired by environmental obstacles or errant button presses that go undetected by the researchers, adding additional sources of potential error. VR owners are also likely to differ from lab participants along various demographic dimensions, including age, ethnicity, and education [10].

This study presents a remote replication of lab-based research on the effects of two teleporting interfaces on triangle completion [11], allowing for comparison between remotely collected data and previously published lab-based data. Participants in the current experiment performed several trials of a triangle completion task remotely using their own VR equipment. They traversed the outbound path using two interfaces: partially concordant teleporting and discordant teleporting. The surrounding VE was a warehouse with ample landmarks placed around the perimeter. Path size was either small or large, and VE size was either small or large.

The hypothesis was that pointing responses would mirror those reported in the lab-based version of this study [11]. In terms of specific expected results, discordant teleporting should lead to worse triangle completion performance compared to partially concordant teleporting, due to differences in the availability of rotational self-motion cues. This was one of the primary conclusions from the lab study, and replication would be evidenced by greater absolute pointing errors when using the discordant compared to partially concordant interface.

Another expected result is that the difference in triangle completion error between the two teleporting interfaces should depend on the proximity of the landmarks, located at the perimeter of the VE, to the traveled path. The difference between the two interfaces should be minimized when navigating large triangles within the small VE, because this specific combination of conditions brings the navigator near the environmental cues in the periphery of the VE, thereby allowing for greater reliance on piloting to complete the task. This was another primary conclusion from the lab study [11].

There are reasons to expect better overall performance from the remote sample (e.g., gaming experience and familiarity with the VR equipment), as well as reasons to expect worse overall performance from the remote sample (e.g., distractions in their home environment). Therefore, comparison of overall errors between the remote sample and

the lab-based sample was considered exploratory.

The research method closely followed the lab-based version of the study [11]. The primary difference between this study and the lab-based version is that participation occurred remotely using the participant's own HMD. However, other differences necessarily occurred. Whereas lab-based participants used the HTC Vive, remote participants used one of several compatible tethered HMDs. Furthermore, the number of trials in the remote study was reduced (from 96 to 48) to prevent disengagement in the absence of a supervising researcher. Reducing the trial numbers also reduced the granularity of the path angles, although the range of path angles was retained. Additionally, blocking of the independent variables was changed: remote participants used one interface for the first half of trials and the other for the second half of trials, in contrast to lab participants who switched between interfaces multiple times. This decision was made to reduce interference and confusion when switching between interfaces, and is justified by the lack of order effects in the lab-based study. Lastly, whereas lab participants were instructed by a trained researcher, online participants were instructed through a short video followed by instruction within the VE, including adaptive feedback that was triggered by specific user errors (e.g., incorrect button presses). None of these changes were expected to impact the primary findings of the lab-based study.

The target sample size was 40 participants with complete data (i.e., with triangle completion data corresponding to each combination of interface, path size, and VE size). To determine sample size, data from 32 participants in the original lab-based study (downloaded from the Open Science Framework: https://osf.io/83vty/) were downsampled by removing half of the data points to simulate the change from 12 trials per condition in the original study to 6 trials per condition in the current study. Next, the analyses reported in the original paper were run using the downsampled data. This process was conducted five times, each time with a newly generated downsampled data set, to ensure that the impact of reducing the trial numbers was fully understood. Of the significant results reported in the original study, the one with the smallest effect size changed from a η_p^2 of .14 using the full data to .10 using the downsampled data (using the smallest value obtained from the five repetitions). Next, a power analysis (G*Power v3.1) was conducted with the following parameters: repeated measures ANOVA, η_p^2 = .10, alpha = .05, minimum power needed to detect an effect = .95, and correlation between repeated measures = .3 (calculated from the original data). This analysis produced a target sample size of 29 participants. Out of concern that distraction during unsupervised research could reduce data quality, the target sample size was increased to 40.

4 METHOD

4.1 Participants

Forty-two participants (37 men, 4 women, 1 declined to state) were recruited through a combination of online work sites (Amazon's Mechanical Turk and Prolific) and social media advertisements. Participants were paid $10 for completion of the study. To be eligible, participants had to be 18 years or older. They also had to have one of the following compatible HMDs connected to SteamVR: Oculus Rift, Oculus Rift S, Oculus Quest (connected via Oculus Link), HTC Vive, HTC Vive Pro, or Valve Index.

4.2 Stimuli

The small VE was 37.5 meters by 37.5 meters and the large VE was 75 meters by 75 meters (see Figure 1 for perspective views and Figure 2 for a schematic overhead view). Both VEs presented an open warehouse with landmarks such as shipping containers, ladders, boxes, and cargo doors, lining the perimeter.

Paths used for the triangle completion task were marked by a sequence of three posts, each 1 meter tall (see Figure 1). A green post marked the path origin, a yellow post marked the end of the first path leg, and a red post marked the end of the second path leg. All three posts had a small, light blue arrow at their base indicating the direction of the next post in the sequence, which was necessary in order to select the intended orientation when using the discordant teleporting interface.
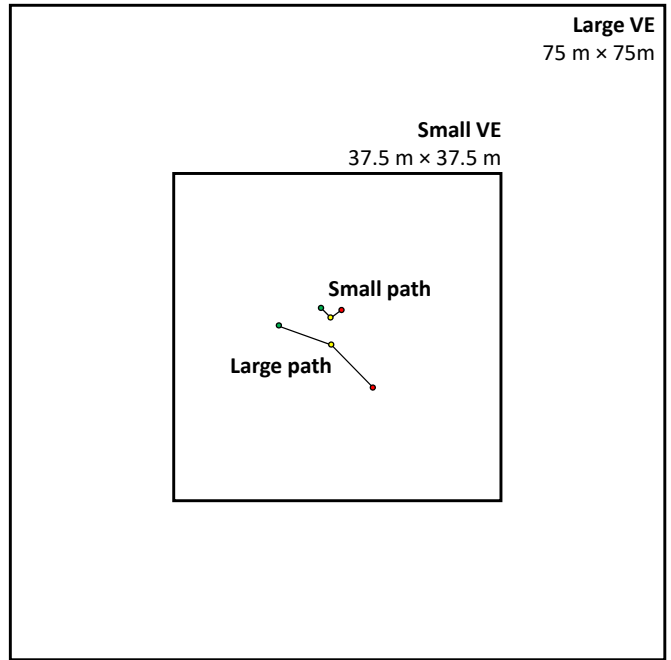

Fig. 2. Overhead view showing the dimensions of the small VE and large VE in relation to sample triangles corresponding to small and large paths.

The arrow at the base of the red post pointed in the same direction as the arrow on the yellow post. The arrows were not necessary for using the partially concordant teleporting interface, but were included for consistency between the two interfaces.

One hand-held controller was used to perform the task. When using the partially concordant teleporting interface, the participant pressed and held the thumbstick button (Valve and Oculus remotes) or trackpad button (HTC remotes) while positioning a white ring (see Figure 1, top row) to select their intended position. Thumbstick / trackpad rotations had no effect when using this interface. Upon releasing the button, the participant was immediately teleported to the selected location without any self-motion cues. Rotations were controlled by rotating the head, and were accompanied by visual and body-based self-motion cues. When using the discordant teleporting interface, the participant pressed and rotated the thumbstick / trackpad to position and orient a magenta ring (see Figure 1, bottom row) in order to select their intended position and orientation. Upon release of the button, the participant was instantly teleported to that position and orientation without any self-motion cues. Both interfaces snapped to the position (and orientation when using the discordant interface) of the post to prevent outbound path errors.

Adaptive feedback was included to facilitate participant training on the task and controls. Incorrect button presses were followed by feedback reminding of the correct button, a long time lapse before teleporting to the next post was followed by a reminder about how to teleport, and a long time lapse before making the pointing response was followed by a reminder about the task goal and which button to press when pointing.

4.3 Design

The independent variables were the teleporting interface (partially concordant or discordant), path size (small or large), and VE size (small or large). Participant sample (remote or lab) was included as a predictor variable in the analyses, but was not technically an independent variable in the design of the current study since it was not manipulated (i.e., this study only collected remote data, which were compared to published lab data [11]).

Teleporting interface was blocked and order was randomized. Within each interface block, path size was blocked and order was randomized. Within each path size block, VE size was blocked and order randomized.

Thus, there were a total of 8 blocks corresponding to every combination of interface, path size, and VE size. Each block contained six triangle completion trials corresponding to turn angles of -135, -90, -45, +45, +90, and +135 degrees, presented in a random sequence. On each trial, the length of each path length was randomly selected from three possible values (small path: 1.52, 1.68, 1.83 meters; large path: 6.1, 6.7, 7.3 meters). Sample paths are shown in Figure 2.

4.4 Procedure

After completing the informed consent form, the participant was directed to a website with instructions about how to download and run the Unity VR software as well as how to perform the triangle completion task. The instructions specified that the participant should attempt to remember their start location, traverse the outbound path (i.e., the two legs of the triangle), and then point back to the start location. The participant was asked to watch two videos demonstrating the task with each teleporting interface. The participant then donned their HMD and performed at least one practice trial in a VE consisting of a ground plane textured with a grid pattern. The formal experiment began when the participant felt comfortable with the task and controls.

Videos showing the participant's view of the triangle completion task can be found on the Open Science Framework. Each triangle completion trial began with a green post marking the path origin. The participant teleported to the green post, which disappeared upon arrival. Next, a yellow post appeared, marking the end of the first path leg. The yellow post disappeared upon arrival, and a red post appeared marking the end of the second path leg. Upon arrival at the red post, the participant indicated the remembered location of the green post by positioning a small blue circle using the controller. A button press logged their response and the next trial ensued. After completion of each six-trial block, the participant was told that they could take a short break or continue to the next block. After completion of the VR task, participants completed questionnaires asking about basic demographic information, frequency of VR use, and hours spent playing video game.

The average study completion time was 43 minutes (SD = 15 minutes).

5 ANALYSIS AND INTERPRETATION OF RESULTS

Per a handbook on replicability [19], the goal of replicability is "obtaining consistent results across studies aimed at answering the same scientific question" (p. 72). This source further notes that while there are many approaches to assessing replicability, any such assessment should account for proximity (e.g., closeness of means), and uncertainty (variability in the results measures). To address this, remote data from the current study were compared with previously published lab data [11] in three ways. First, distributions of response locations were compared between remote and lab data by computing a Bhattacharyya coefficient [3], which describes the amount of overlap between two distributions. This was done for each of the eight conditions created by the 2 (interface: partially concordant or discordant) × 2 (VE size: small or large) × 2 (path size: small or large) experimental design.

Second, each dependent variable from the remote sample data was analyzed in a 2 × 2 × 2 repeated measures ANOVA. These analyses parallel analyses from the lab study, and are referred to as remote ANOVAs because they include only remote participant data. For each dependent variable, remote ANOVA results are juxtaposed with lab ANOVA results reported in the lab-based study to identify consistencies and inconsistencies in the conclusions drawn.

Figure 3, which uses iconography instead of numbers to compare remote ANOVA results (described in detail below) with lab ANOVA results, is presented as a tool to encourage holistic evaluation of the consistency across results. Interpretation of the patterns in Figure 3 is reserved for the discussion.

Third, each dependent variable was analyzed in a 2 (sample: remote or lab) by 2 (interface: partially concordant or discordant) by 2 (VE size: small or large) by 2 (path size: small or large) ANOVA. These are referred to as omnibus ANOVAs because they include all primary independent variables as well as the participant sample variable. A significant main effect of sample or an interaction between sample

and another independent variable would indicate that differences exist between data collected from the remote and lab studies.

6 RESULTS

Data from one male participant were removed for absolute errors in excess of three standard deviations above the mean. Of the remaining participants, average age was 28.26 years (SD=8.35). When asked about frequency of VR use, the two most common responses were "Once or twice a week" (n=14) and "Once or twice a month" (n=14), with an average VR session length of 71.49 minutes (SD=39.07). Participants reported playing video games (not necessarily in VR) an average of 29.03 hours per week (SD=14.37), which is consistent with recent research on VR headset owners [10].

6.1 Bhattacharyya coefficients

Bhattacharyya coefficients (BC) [3] were calculated to evaluate the similarity of the 2D response distributions from remote and lab participants. BC can range from 0, indicating no overlap, to 1, indicating perfect overlap. These calculations, presented in Table 1, were conducted separately for each of the eight conditions in the experimental design. For each condition, the pointing response distributions compared using BC were the x, y locations of participants' responses relative to the location of the green post, set at the origin of the x, y plane. Because the correct direction and distance of the path origin varied from trial to trial, with variations in the size, angle, and orientation of the triangle, responses were calculated relative to the correct direction and distance of the path origin from the path terminus. Furthermore, because the turn angles in triangle completion trials were clockwise or counterclockwise, responses for triangles with a counterclockwise turn were flipped such that negative angles reflected under-rotation when turning toward the path origin (assuming participants rotated the shorter direction).

BC comparing remote to lab responses indicated high similarity ($\geq$ .923) in seven of the eight conditions. BC indicated only moderate similarity (.767) in the condition involving the partially concordant teleporting interface, small VE, and small path. Closer inspection of responses in that condition indicated that remote and lab samples produced quite similar response angles (BC = .960) and response distances (BC = .979) when compared separately. However, the conjunction of response angles and distances revealed that in this condition only, lab and remote participants both had subtle but opposite correlations between their response distance and response angle, with lab participants demonstrating a weak positive correlation, and remote participants having a weak negative correlation. These analyses are presented in more detail on the Open Science Framework.

6.2 Absolute Distance Error

Absolute distance error is the absolute distance between the location of the participant's pointing response and the location of the path origin. Results are displayed in Figure 4. The remote ANOVA, which includes remote participant data only, revealed significant main effects of interface, $F(1,41) = 30.314$, $p <.001$, $\eta_p^2 = .425$, VE size, $F(1,41) = 16.610$, $p <.001$, $\eta_p^2 = .288$, and path size, $F(1,41) = 206.568$, $p <.001$, $\eta_p^2 = .834$, as well as a significant interaction between VE size and path

Table 1. BC calculations comparing the distributions of 2D pointing responses made by lab and remote participants, by condition.

Condition	Bhattacharyya Coefficient
Partially concordant, small VE, small path	0.767
Partially concordant, small VE, large path	0.941
Partially concordant, large VE, small path	0.954
Partially concordant, large VE, large path	0.923
Discordant, small VE, small path	0.968
Discordant, small VE, large path	0.966
Discordant, large VE, small path	0.965
Discordant, large VE, large path	0.958

ANOVA effect	Abs. distance error		Abs. angular error		Signed angular error		Absolute axial error		Signed axial error		Response time	
	Lab	Remote	Lab	Remote	Lab	Remote	Lab	Remote	Lab	Remote	Lab	Remote
Interface	✓ ●	✓ ●	✓ ●	✓ ●	✓ ●	✓ ●	✓ ●	✓ ●	✗ ○	✗ ○	✓ ●	✓ ●
VE size	✓ ●	✓ ●	✓ ●	✓ ◐	✓ ●	✗ ○	✓ ●	✓ ●	✗ –	✗ –	✗ ○	✗ ○
Path size	✓ ●	✓ ●	✓ ●	✓ ●	✗ –	✗ –	✗ ○	✗ ○	✓ ●	✓ ●	✓ ●	✓ ●
Interface*VE size	✓ ●	✗ ○	✓ ◐	✗ –	✗ –	✓ ◐	✗ ○	✗ ○	✗ ○	✗ ◐	✗ ○	✗ ○
Interface*Path size	✓ ◐	✗ ○	✓ ●	✗ ○	✗ –	✗ –	✗ –	✗ –	✗ ○	✓ ●	✗ ○	✗ ○
VE size*Path size	✓ ●	✓ ●	✓ ●	✓ ◐	✓ ●	✗ –	✗ –	✗ ◐	✗ ◐	✗ ○	✗ –	✗ ◐
Interface*VE size*Path size	✗ ○	✗ –	✗ ○	✗ –	✗ –	✗ –	✗ –	✗ ○	✗ –	✗ –	✗ ○	✗ ○

Fig. 3. Comparison of significant (green check) and non-significant (red X) ANOVA results from the lab participant sample (n=32) [11] and the remote participant sample (i.e., "remote ANOVA" from the current study, n=41). Light gray boxes indicate consistent results and medium gray boxes indicate inconsistent results in terms of statistical significance. Effect sizes are represented by symbols corresponding to partial eta squared values: dash indicates a negligible effect ($\eta_p^2 < .01$), open circle indicates a small effect ($.01 \leq \eta_p^2 < .06$), half-filled circle indicates a medium effect ($.06 \leq \eta_p^2 < .14$), and filled circle indicates a large effect ($\eta_p^2 \geq .14$).

size, $F(1,41) = 11.398$, $p = .002$, $\eta_p^2 = .218$. These findings were also reported in the lab-based sample. However, other significant results from the lab-based data did not occur in the remote data. Specifically, the lab-based data produced significant interactions between interface and VE size as well as interface and path size, and these interactions were not significant in the remote data. The consistent and inconsistent ANOVA results are highlighted in Figure 3.

The omnibus ANOVA, which includes data from remote and lab samples, revealed significant main effects of interface, $F(1,72) = 75.459$, $p < .001$, $\eta_p^2 = .512$, VE size, $F(1,72) = 53.529$, $p < .001$, $\eta_p^2 = .426$, and path size, $F(1,72) = 328.356$, $p < .001$, $\eta_p^2 = .820$, as well as interactions between interface and VE size, $F(1,72) = 4.257$, $p = .043$, $\eta_p^2 = .056$, interface and path size, $F(1,72) = 6.460$, $p = .013$, $\eta_p^2 = .082$, and VE size and path size, $F(1,72) = 47.277$, $p < .001$, $\eta_p^2 = .396$. The sample variable (remote vs. lab) was not significant, $F(1,72) = 2.774$, $p = .10$, $\eta_p^2 = .037$, nor were any interactions involving sample.

6.3 Absolute Angular Error

Absolute angular error is the absolute angular distance between the direction of the participant's pointing response and the direction of the path origin. Results are displayed in Figure 5. The remote ANOVA,
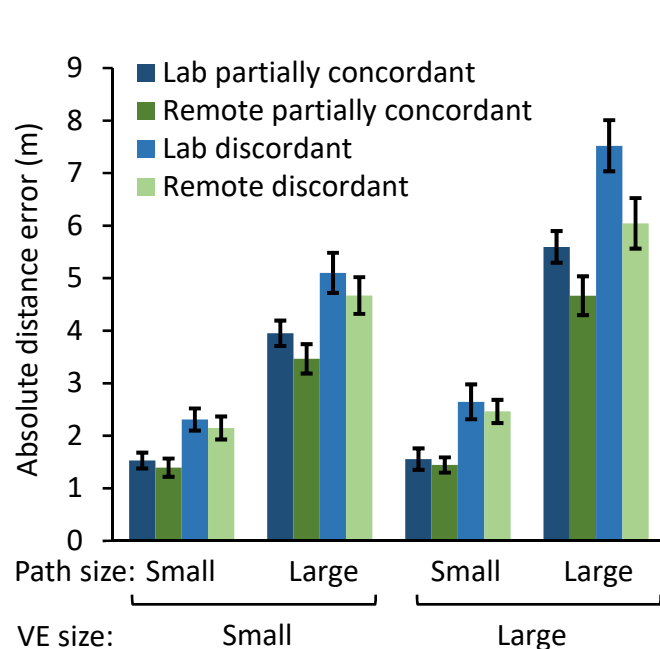
which includes remote participant data only, revealed significant main effects of interface, $F(1,41) = 30.142$, $p < .001$, $\eta_p^2 = .424$, VE size, $F(1,41) = 4.297$, $p = .044$, $\eta_p^2 = .095$, and path size, $F(1,41) = 20.537$, $p < .001$, $\eta_p^2 = .334$, as well as an interaction between VE size and path size, $F(1,41) = 6.077$, $p < .018$, $\eta_p^2 = .129$. These findings were also reported in the lab-based sample. However, other significant results from the lab-based data did not occur in the remote data. Specifically, the lab-based data produced significant interactions between interface and VE size as well as interface and path size, and these interactions were not significant in the remote data.

The omnibus ANOVA revealed significant main effects of interface, $F(1,72) = 81.086$, $p < .001$, $\eta_p^2 = .530$, warehouse size, $F(1,72) = 19.651$, $p < .001$, $\eta_p^2 = .214$, and path size, $F(1,72) = 38.419$, $p < .001$, $\eta_p^2 = .348$, as well as interactions between interface and path size, $F(1,72) = 6.912$, $p = .010$, $\eta_p^2 = .088$, as well as VE size and path size, $F(1,72) = 18.774$, $p < .001$, $\eta_p^2 = .207$. The sample variable (remote vs. lab) was not significant, $F(1,72) = .491$, $p = .486$, $\eta_p^2 = .007$, nor were any interactions involving sample.

Fig. 4. Average absolute distance error as a function of path size, VE size, and teleporting interface. Error bars represent ± 1 SE of the mean.
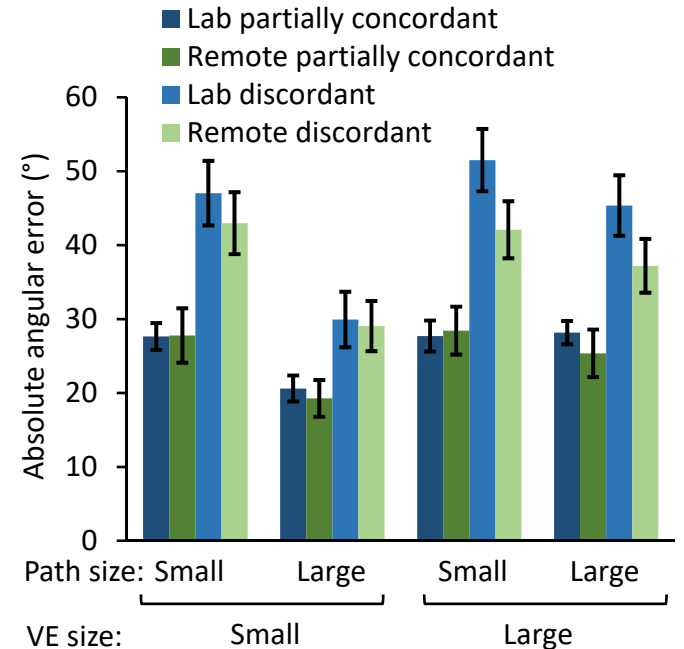
Fig. 5. Average absolute angular error as a function of path size, VE size, and teleporting interface. Error bars represent ± 1 SE of the mean.
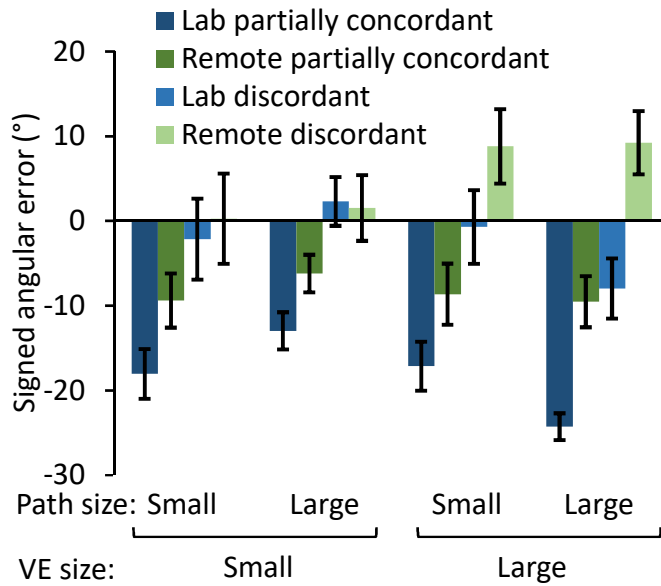
Fig. 6. Average signed angular error as a function of path size, VE size, and teleporting interface. Error bars represent $\pm$ 1 SE of the mean.

Fig. 7. Average absolute axial error as a function of path size, VE size, and teleporting interface. Error bars represent $\pm$ 1 SE of the mean.

6.4 Signed Angular Error

Signed angular error is the signed angular distance between the direction of the participant's pointing response and the direction of the path origin. Data from trials with counter-clockwise turns (i.e., the path turn at the yellow post) were flipped in order to align signed errors with those from trials with clockwise turns. Following this transformation, positive errors reflect over-rotation when turning to face the path origin (assuming the participant turned the shorter direction when doing so). Results are displayed in Figure 6.

The remote ANOVA revealed a significant main effect of interface, $F(1,41) = 14.735$, $p < .001$, $\eta_p^2 = .264$, which was also reported in the lab-based sample, as well as an interaction between interface and VE size, $F(1,41) = 5.939$, $p = .019$, $\eta_p^2 = .127$, which did not occur in the lab-based sample. The lab-based study additionally reported a significant interaction between VE size and path size that was not significant in the remote sample.

The omnibus ANOVA revealed significant main effects of sample, $F(1,72) = 9.950$, $p = .002$, $\eta_p^2 = .121$, and interface, $F(1,72) = 45.197$, $p < .001$, $\eta_p^2 = .386$, as well as interactions between sample and VE size, $F(1,41) = 6.103$, $p = .016$, $\eta_p^2 = .078$, and between VE size and path size, $F(1,41) = 5.747$, $p = .019$, $\eta_p^2 = .074$. The main effect of sample reflected the fact that remote participant responses showed less negative angular error, on average, than did lab participant responses (lab sample: $M = -10.127$, $SE = 2.0$; remote sample, $M = -1.752$, $SE = 1.746$). In other words, remote participants had a reduced tendency to under-rotate when turning to face and point to the path origin. The interaction between sample and VE size occurred because lab participants under-rotated more in the large VE ($M = -12.534$, $SE = 2.321$) compared to the small VE ($M = -7.719$, $SE = 2.401$), whereas remote participants under-rotated less in the large VE ($M = -.044$, $SE = 2.026$) compared to the small VE ($M = -3.460$, $SE = 2.096$).

6.5 Absolute Axial Error

Absolute axial error is the absolute difference between the distance of the pointing response and the distance of the path origin divided by the distance of the path origin. Results are displayed in Figure 7. The remote ANOVA revealed significant main effects of interface, $F(1,41) = 36.097$, $p < .001$, $\eta_p^2 = .468$, and VE size, $F(1,41) = 8.402$, $p = .006$, $\eta_p^2 = .170$. These results are consistent with those reported in the lab-based sample.

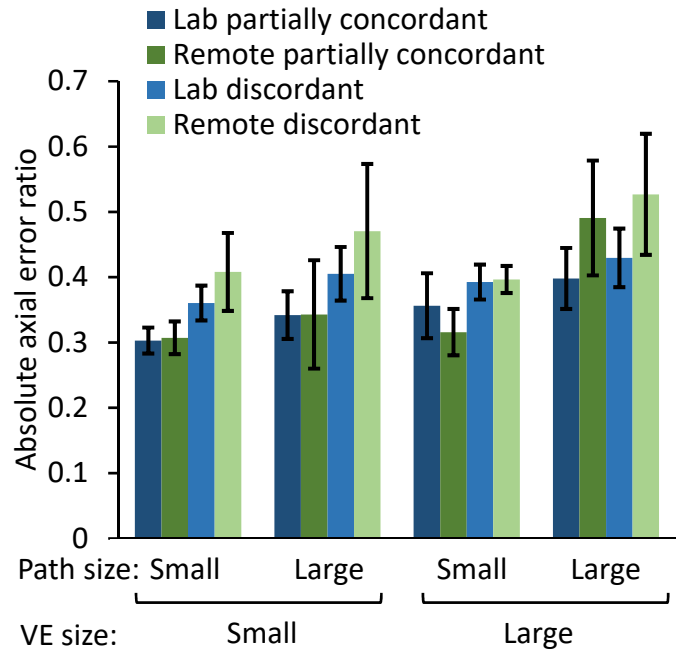The omnibus ANOVA revealed significant main effects of interface,

$F(1,72) = 35.845$, $p < .001$, $\eta_p^2 = .332$, and VE size, $F(1,72) = 13.069$, $p = .001$, $\eta_p^2 = .154$. The sample variable was not significant, $F(1,72) = .301$, $p = .585$, $\eta_p^2 = .004$, nor were any interactions involving sample.

6.6 Signed Axial Error

Signed axial error is the distance to the pointing response minus the distance to the path origin divided by the distance to the path origin. Negative axial error reflects responses that were too short and positive axial error reflects responses that were too long. Results are displayed in Figure 8. The remote ANOVA revealed a significant main effect of path size, $F(1,41) = 19.823$, $p < .001$, $\eta_p^2 = .326$, as well as a significant interaction between interface and path size, $F(1,41) = 9.469$, $p = .004$, $\eta_p^2 = .188$. These main effect of path size is consistent with results from the lab-based sample, but the interaction between interface and path size is unique to the remote sample. This interaction reflected the fact that small path errors were larger with the discordant interface compared to the partially concordant interface, whereas large path errors showed the opposite trend.

The omnibus ANOVA revealed a significant main effect of path length, $F(1,72) = 32.367$, $p < .001$, $\eta_p^2 = .310$, as well as interactions between interface and VE size, $F(1,72) = 6.989$, $p = .010$, $\eta_p^2 = .088$, interface and path length, $F(1,72) = 8.794$, $p = .004$, $\eta_p^2 = .109$, and a three-way interaction between sample, VE size, and path length, $F(1,72) = 5.121$, $p = .027$, $\eta_p^2 = .066$. The three-way interaction was the only significant effect involving the sample variable. The main effect of sample was not significant, $F(1,72) = .064$, $p = .80$, $\eta_p^2 = .001$. Closer inspection of the data indicates that remote and lab participants produced similar errors in the small VE regardless of path size. However, errors in the large VE indicate that remote participants, compared to lab participants, produced longer responses on small paths and shorter responses on large paths.

6.7 Response Time

Response time is the difference between when the participant arrived at the end of the path and when the pointing response was logged. Results are displayed in Figure 9. The remote ANOVA revealed significant main effects of interface, $F(1,41) = 13.709$, $p = .001$, $\eta_p^2 = .251$, and path size, $F(1,41) = 25.860$, $p < .001$, $\eta_p^2 = .387$. These results are
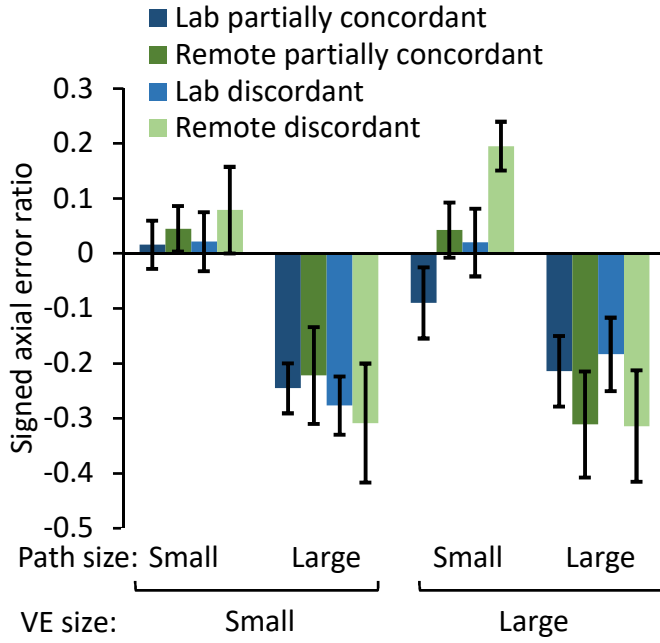
Fig. 8. Average signed axial error as a function of path size, VE size, and teleporting interface. Error bars represent ± 1 SE of the mean.

consistent with those reported in the lab-based sample.

The omnibus ANOVA revealed significant main effects of interface, $F(1,72) = 28.142$, $p < .001$, $\eta_p^2 = .281$, and path size, $F(1,72) = 35.864$, $p < .001$, $\eta_p^2 = .332$. The sample variable was not significant, $F(1,72) = .695$, $p = .407$, $\eta_p^2 = .01$, nor were any interactions involving sample.

7 DISCUSSION

Triangle completion data collected from remote participants were analyzed and compared with previously published data collected from lab participants [11]. This comparison between remote and lab data
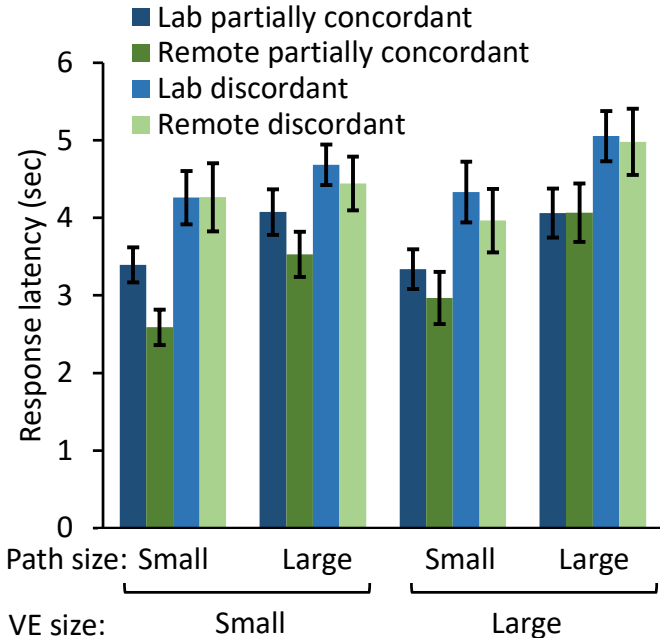

Fig. 9. Average response time as a function of path size, VE size, and teleporting interface. Error bars represent ± 1 SE of the mean.

took three forms. First, Bhattacharyya coefficients were calculated to evaluate the similarity between the response distributions produced by remote and lab participants. This was calculated separately for each of the eight combinations of interface, VE size, and path size. Seven of the eight conditions indicated high similarity of responses. One condition (partially concordant teleporting, small VE, and small path) was lower than the others but still indicated considerable overlap. Although it is unclear why the two participant samples diverged somewhat in that particular condition, the results overall indicate close alignment of responses between remote and lab participants.

Second, remote data were analyzed in a manner parallel to that reported in the lab-based study. Results from these analyses of the remote data were juxtaposed with those from the lab data to identify consistencies and inconsistencies in the conclusions drawn. These analyses provided mixed evidence for replication, and are discussed in detail below.

Third, data were directly compared in an omnibus ANOVA with sample (remote vs. lab) as one of several independent variables. These analyses provided good evidence in favor of replication, and are also discussed in detail below.

Parallel analysis of remote and lab data indicated that several, but not all, of the findings were consistent across the two samples (see visualization in Figure 3). Of the 18 main effects tested (three main effects across six dependent variables), 17 were consistent across the two samples. Of the 24 interactions tested, 17 were consistent across samples. Following this approach to comparing the two data sets, the conclusions drawn from the remote and lab samples shared many similarities but also diverged in potentially important ways.

One way to unpack the similarities and differences revealed by the parallel analyses is to evaluate whether major conclusions originally reported using the lab data [11] were supported by the remote data. One of the primary conclusions based on the lab data was that the discordant teleporting interface, which lacks the rotational self-motion cues available in the partially concordant teleporting interface, led to worse triangle completion performance. This conclusion was based on the finding that absolute distance error and absolute angular error were higher with the discordant teleporting interface compared to the partially concordant teleporting interface. Similar analyses with the remote data supported the same conclusion, which is further bolstered by research demonstrating the importance of rotational self-motion cues for navigation generally [7, 13, 14, 24, 25].

Another conclusion from the lab study [11] was that the importance of rotational self-motion cues (i.e., the difference between partially concordant and discordant teleporting interfaces) depended on the proximity of piloting cues. Specifically, the difference in error (absolute distance error as well as absolute angular error) between the two interfaces was reduced when navigating large triangles within the small VE because that specific combination brought the participant closer to the piloting cues lining the periphery of the VE. Reliability of piloting cues depends on their distance from the navigator (i.e., nearer landmarks provide a more reliable estimate of self-location), and more reliable piloting cues allowed participants to place greater weight on piloting than on path integration. This conclusion was supported by significant interactions in the lab data between interface and VE size, interface and path size, as well as follow-up analyses that more closely examined those interactions. This finding is also consistent with navigation research indicating that humans are capable of assigning cue weights based on relative cue reliability [5, 18, 27]. However, the expected interactions were not significant in the remote data. Closer inspection of absolute angular errors reveals that remote and lab participants performed similarly with the partially concordant interface, but that remote participants performed slightly better with the discordant interface when piloting cues were far away (i.e., the small path in the small VE as well as small and large paths in the large VE), which may have reduced the benefit of nearby piloting cues (i.e., the large path in the small VE). This difference between remote and lab participants might reflect differences in VR experience. Experienced VR users (i.e., participants in the remote sample) may be better able to update self-location using a challenging locomotion interface such as the discordant teleporting interface.

The above description addresses differences between remote and lab ANOVAs in absolute distance error and absolute angular error. What might explain other differences visualized in Figure 3, specifically differences in the signed angular error and signed axial error data? Regarding signed angular error, the original study reported that under-rotation toward the path origin was more pronounced in the large VE compared to the small VE, especially when navigating large paths. This effect did not replicate in the remote sample. The original study made no predictions about signed angular error, nor did the original study provide a compelling explanation for these effects, so it is unclear how to interpret the failure to replicate this effect in the current study. Regarding signed axial errors, the remote data produced shorter response distances with the discordant interface compared to the partially concordant interface, and this trend was exaggerated on large paths compared to small paths. This is also an non-predicted interaction, and one that defies clear explanation. We acknowledge that these discrepant findings are partial evidence against replication. However, it is unclear whether these discrepancies are meaningful differences because the results in question do not follow from theory nor practical inference.

The third approach to comparing responses from remote and lab participants involved joint analysis of remote and lab data in omnibus ANOVAs. Across six dependent variables, there was one significant main effect of the sample variable. Of the 42 possible interactions involving the sample variable (seven possible interactions in each ANOVA across six dependent variables), two were statistically significant. Of course, lack of statistical significance should not be interpreted evidence of equivalence. However, these results do indicate that few major differences exist between triangle completion performance by remote participants and lab participants.

The differences between remote and lab samples were found in the signed angular errors and signed axial errors, and they appear to reflect subtle differences in biases toward under-rotation as well as biases in pointing distance when pointing toward the path origin. Whereas responses by remote participants were slightly less biased in terms of signed angular error, they were slightly more biased in terms of signed axial error. We see no obvious explanation for these differences, but future work may produce a more principled explanation.

To summarize, Bhattacharyya coefficients calculated for each of eight conditions indicated a high degree of similarity between remote and lab participant responses in seven conditions and moderate similarity in one condition. Parallel analysis of remote and lab data indicated many similarities and also multiple differences. In one case, the difference between remote and lab data resulted in different decisions about an a priori hypothesis that piloting cues exert greater influence when they are closer to the navigator. Direct comparison of remote and lab data within the same model uncovered few significant differences between the groups. The differences that did occur were found in response biases, and lack clear explanation.

8 LIMITATIONS

Although many primary findings from this remote, unsupervised study on locomotion interfaces replicated those from a previously published lab-based study [11], several effects did not replicate (see Figure 3). There is no clear explanation for the divergent results. Possible explanations include different participant characteristics (e.g., age or video game experience), different instructions provided to in-person and remote participants, different equipment used by in-person and remote participants, and different characteristics of the testing environment (e.g., environmental distraction). The vast number of differences between the remote and lab-based studies make it exceedingly difficult to track down the root causes of divergent results, and this problem is likely to characterize future attempts to replicate lab-based research in a remote, unsupervised setting. As is often true in research, successes are easier to interpret than failures.

It is encouraging that several primary findings from the lab-based study did replicate under remote, unsupervised conditions. However, it cannot be taken for granted that other findings established under lab-based conditions will replicate under remote, unsupervised conditions. Other tasks may be more sensitive to environmental influences such as external distraction, or to demographic differences between the populations from which participants are recruited. A cautious approach is warranted when researchers hope to extend theories and methods established under lab conditions by conducting remote research.

More generally, there are limitations to remote, unsupervised research that will make some VR research questions impossible or unfeasible to test remotely. For example, research requiring specialized equipment (e.g., a treadmill device or eye tracking equipment) is not likely to be feasible as a remote study. Research involving more complex tasks often involves a researcher who actively observes and occasionally interacts with the participant, and this type of research would be far more difficult to implement in a remote, unsupervised setting. Other limitations to conducting remote, unsupervised research include considerations of participant safety, as well as the amount of researcher effort required to create an application that supplants the guidance provided by a live, in-person researcher.

9 CONCLUSIONS

This study evaluated whether findings from a lab-based study on locomotion interfaces for VR [11] would replicate under remote, unsupervised testing conditions. Many of the lab results did replicate under remote conditions, despite differences in equipment, experimenter supervision, testing environment, and participant sample. Most notably, triangle completion performance was better when rotational self-motion cues were available. Yet, several differences were found between the remote and lab data. Some of those differences involved effects that are not easily explained, such as non-predicted interactions between specific variables. It is possible that those effects would not replicate even under identical lab conditions. Perhaps the most important difference is that the lab results indicated that nearby piloting cues lessen the difference between the discordant and partially concordant teleporting interfaces, but the remote data did not support this conclusion. One possible explanation is that experienced VR users in the remote sample were better able to manage the challenges of the discordant teleporting interface. Consistent with research showing that navigators adjust relative cue weights based on cue reliability [5, 18, 27], it is possible that VR experience increases the reliability of information gained about self-position from the discordant interface, resulting in a difference in relative cue weights for those with and without VR experience.

Researchers hoping to produce findings that generalize between lab participants and remote participants should proceed cautiously. Data collected under remote, unsupervised conditions can be sufficiently sensitive to detect predicted differences, such as the difference between two interfaces that vary in available self-motion cues. However, close comparison to data collected in a supervised setting may also reveal important differences. These differences could be caused by differences in the research setting, as well as differences in participant characteristics [10]. Researchers should strongly consider replicating their research under lab and remote conditions to compare findings before further evaluating research questions using entirely remote data.

REFERENCES

[1] G. L. Allen, K. C. Kirasic, M. A. Rashotte, and D. B. M. Haun. Aging and path integration skill: Kinesthetic and vestibular contributions to wayfinding. *Perception & Psychophysics*, 66:170–179, 2004. doi: 10. 3758/BF03194870

[2] M. D. Barrera Machuca and W. Stuerzlinger. *The Effect of Stereo Display Deficiencies on Virtual Hand Pointing*, p. 1–14. ACM, New York, 2019.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta math. Soc.*, 35:99–109, 1943.

[4] D. A. Bowman, D. Koller, and L. F. Hodges. Travel in immersive virtual environments: An evaluation of viewpoint motion control techniques. In *Proc. Virtual Reality Annual International Symposium*, pp. 45–52. IEEE Computer Society, Washington, D.C., 1997. doi: 10.1109/VRAIS.1997.583043

[5] X. Chen, T. P. McNamara, J. W. Kelly, and T. Wolbers. Cue combination in human spatial navigation. *Cognitive Psychology*, 95:105–144, 2017. doi: 10.1016/j.cogpsych.2017.04.003

[6] L. A. Cherep, J. W. Kelly, A. J. Miller, A. F. Lim, and S. B. Gilbert. Individual differences in teleporting through virtual environments. *Journal of Experimental Psychology: Applied*, 2021. doi: 10.31234/osf.io/b6cyd

[7] L. A. Cherep, A. F. Lim, J. W. Kelly, D. Acharya, A. Velasco, E. Bustamante, A. G. Ostrander, and S. B. Gilbert. Spatial cognitive implications of teleporting through virtual environments. *Journal of Experimental Psychology: Applied*, 26(3):480–492, 2020. doi: 10.1037/xap0000263

[8] T. Hartig, M. Mang, and G. W. Evans. Restorative effects of natural environment experiences. *Environment and Behavior*, 23(1):3–26, 1991. doi: 10.1177/0013916591231001

[9] B. Huber and K. Z. Gajos. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *PLoS ONE*, 15(1):ee0227629, 2020. doi: 10.1371/journal.pone.0227629

[10] J. W. Kelly, L. A. Cherep, A. F. Lim, T. Doty, and S. B. Gilbert. Who are virtual reality headset owners? a survey and comparison of headset owners and non-owners. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 687–694, 2021. doi: 10.1109/VR50410.2021.00095

[11] J. W. Kelly, A. G. Ostrander, A. F. Lim, L. A. Cherep, and S. B. Gilbert. Teleporting through virtual environments: Effects of path scale and environment scale on spatial updating. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1841–1850, 2020. doi: 10.1109/TVCG.2020.2973051

[12] J. W. Kelly, J. Terrill, M. Zimmerman, T. A. Doty, L. A. Cherep, M. T. Hoover, N. R. Powell, O. J. Perrin, and S. B. Gilbert. Boundaries facilitate spatial orientation in virtual environments. In *Symposium on Spatial User Interaction*, SUI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3485279.3488284

[13] R. L. Klatzky, J. M. Loomis, A. C. Beall, S. S. Chance, and R. G. Golledge. Spatial updating of self-position and orientation during real, imagined, and virtual locomotion. *Psychological Science*, 9:293–298, 1998. doi: 10.1111/1467-9280.00058

[14] A. F. Lim, J. W. Kelly, N. C. Sepich, L. A. Cherep, G. C. Freed, and S. B. Gilbert. Rotational self-motion cues improve spatial learning when teleporting in virtual environments. In *Symposium on Spatial User Interaction*, SUI '20. ACM, New York, 2020. doi: 10.1145/3385959.3418443

[15] X. Ma, M. Cackett, L. Park, E. Chien, and M. Naaman. Web-based vr experiments powered by the crowd. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, p. 33–43. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018. doi: 10.1145/3178876.3186034

[16] S. Milgram, L. Bickman, and L. Berkowitz. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, 13(2):79–82, 1969. doi: 10.1037/h0028070

[17] A. Mottelson, G. B. Petersen, K. Lilija, and G. Makransky. Conducting unsupervised virtual reality user studies online. *Frontiers in Virtual Reality*, 2:66, 2021. doi: 10.3389/frvir.2021.681482

[18] M. Nardini, P. Jones, R. Bedford, and O. Braddick. Development of cue integration in human navigation. *Current Biology*, 18(9):689–693, 2008. doi: 10.1016/j.cub.2008.04.021

[19] National Academies of Sciences, Engineering, and Medicine. Replicability. In *Reproducability and Replicability in Science*, pp. 71–104. National Academies Press, 2019. doi: 10.17226/25303

[20] P. M. Newman and T. P. McNamara. A comparison of methods of assessing cue combination during navigation. *Behavior Research Methods*, 53(1):390–398, 2021. doi: 10.3758/s13428-020-01451-y

[21] T. Nguyen-Vo, B. E. Riecke, W. Stuerzlinger, D.-M. Pham, and E. Kruijff. Naviboard and navichair: Limited translation combined with full rotation for efficient virtual locomotion. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):165–177, 2021. doi: 10.1109/TVCG.2019.2935730

[22] T. C. Peck, L. E. Sockol, and S. M. Hancock. Mind the gap: The underrepresentation of female participants and authors in virtual reality research. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1945–1954, 2020.

[23] G. B. Petersen, A. Mottelson, and G. Makransky. *Pedagogical Agents in Educational VR: An in the Wild Study*. ACM, New York, 2021.

[24] J. J. Rieser. Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(6):1157–1165, 1989. doi: 10.1037/0278-7393.15.6.1157

[25] R. A. Ruddle and S. Lessels. For efficient navigational search, humans require full physical movement, but not a rich visual scene. *Psychological Science*, 17(6):460–465, 1998. doi: 10.1111/j.1467-9280.2006.01728.x

[26] D. Saffo, S. Di Bartolomeo, C. Yildirim, and C. Dunne. *Remote and Collaborative Virtual Reality Experiments via Social VR Platforms*. ACM, New York, 2021.

[27] L. A. Sjolund, J. W. Kelly, and T. P. McNamara. Optimal combination of environmental cues and path integration during navigation. *Memory and Cognition*, 46(1):89–99, 2018. doi: 10.3758/s13421-017-0747-7

[28] A. Steed, S. Frlston, M. M. Lopez, J. Drummond, Y. Pan, and D. Swapp. An 'in the wild' experiment on presence and embodiment using consumer virtual reality equipment. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1406–1414, 2016. doi: 10.1109/TVCG.2016.2518135

[29] A. Steed, F. R. Ortega, A. S. Williams, E. Kruijff, W. Stuerzlinger, A. U. Batmaz, A. S. Won, E. S. Rosenberg, A. L. Simeone, and A. Hayes. Evaluating immersive experiences during covid-19 and beyond. *Interactions*, 27(4):62–67, July 2020. doi: 10.1145/3406098

[30] A. Tang, M. Tory, B. Po, P. Neumann, and S. Carpendale. *Collaborative Coupling over Tabletop Displays*, p. 1181–1190. ACM, New York, 2006.

[31] R. S. Ulrich, R. F. Simons, B. D. Losito, E. Fiorito, M. A. Miles, and M. Zelson. Stress recovery during exposure to natural and urban environments. *Journal of Environmental Psychology*, 11(3):201–230, 1991. doi: 10.1016/S0272-4944(05)80184-7

[32] N. Yee and J. Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33(3):271–290, 07 2007. doi: 10.1111/j.1468-2958.2007.00299.x

[33] L. Zhang and W. Mou. Piloting systems reset path integration systems during position estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(3):472–491, 2017. doi: 10.1037/xlm0000324

[34] J. Zhao, M. Simpson, P. Sajjadi, J. O. Wallgrün, P. Li, M. M. Bagher, D. Oprean, L. Padilla, and A. Klippel. Crowdxr - pitfalls and potentials of experiments with remote participants. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 450–459, 2021. doi: 10.1109/ISMAR52148.2021.00062